Osvaldo Gervasi · Beniamino Murgante ·
Chiara Garau · Ana Maria A. C. Rocha ·
M. Noelia Faginas Lago · Ana C. Braga ·
Yeliz Karaca (Eds.)

LNCS 16760

Computational Science and Its Applications – ICCSA 2026 Workshops

Braga, Portugal, June 30 – July 3, 2026
Proceedings, Part III

3 Part III



Springer

Lecture Notes in Computer Science **16760**

Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*
Wen Gao, *Peking University, Beijing, China*
Bernhard Steffen, *TU Dortmund University, Dortmund, Germany*
Moti Yung, *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

Osvaldo Gervasi · Beniamino Murgante ·
Chiara Garau · Ana Maria A. C. Rocha ·
M. Noelia Faginas Lago · Ana C. Braga ·
Yeliz Karaca
Editors

Computational Science and Its Applications – ICCSA 2026 Workshops

Braga, Portugal, June 30 – July 3, 2026
Proceedings, Part III

Springer

Editors
Osvaldo Gervasi
University of Perugia
Perugia, Perugia, Italy

Chiara Garau
University of Cagliari
Cagliari, Italy

M. Noelia Faginas Lago
University of Perugia
Perugia, Italy

Yeliz Karaca
University of Massachusetts
Worcester, MA, USA

Beniamino Murgante
University of Basilicata
Potenza, Italy

Ana Maria A. C. Rocha
University of Minho
Braga, Portugal

Ana C. Braga
University of Minho
Braga, Portugal

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-032-30520-6 ISBN 978-3-032-30521-3 (eBook)
https://doi.org/10.1007/978-3-032-30521-3

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

These 9 volumes (LNCS volumes 16758–16766) comprise the peer-reviewed papers from the 57 Workshops of the 2026 International Conference on Computational Science and Its Applications (ICCSA 2026), held between June 30 and July 3, 2026, in Braga, Portugal. The peer-reviewed papers from the six main conference tracks are published in a separate set of 3 volumes (LNCS volumes 16767–16769).

The conference was held in a hybrid format, with the vast majority of participants attending in person, hosted by the University of Minho in Braga, Portugal.

We enabled virtual participation for those who did not attend the event in person due to logistical, political, or economic constraints by adopting a technological infrastructure based on open-source software (Jitsi + Riot) and a commercial Cloud infrastructure. ICCSA 2026 was organized by the University of Minho (Portugal), the University of Perugia (Italy), the University of Basilicata (Italy), the University of Cagliari (Italy), Monash University (Australia), and Kyushu Sangyo University (Japan).

With the 2026 edition, ICCSA celebrated its 26th anniversary, marking a new chapter in its longstanding history after its memorable 25th anniversary celebrated in Istanbul, Türkiye. Hosted in Braga, Portugal, the conference found a particularly meaningful setting in one of Europe's oldest and most vibrant academic cities, a place where historical depth, cultural continuity, and contemporary scientific exchange meet. Braga offers an inspiring European context for renewing the international and intercultural dialogue that has always characterized ICCSA, while giving new visibility to the role of computational science in addressing the complex challenges of cities, territories, technologies, and societies.

ICCSA 2026 marked another fruitful and thought-provoking academic event in the International Conference on Computational Science and Its Applications (ICCSA) conference series, previously held in Istanbul, Türkiye (2025), Hanoi, Vietnam (2024), Athens, Greece (2023), Málaga, Spain (2022), Cagliari, Italy (hybrid with few participants in presence in 2021 and completely online in 2020), Saint Petersburg, Russia (2019), Melbourne, Australia (2018), Trieste, Italy (2017), Beijing, China (2016), Banff, Canada (2015), Guimarães, Portugal (2014), Ho Chi Minh City, Vietnam (2013), Salvador, Brazil (2012), Santander, Spain (2011), Fukuoka, Japan (2010), Suwon, South Korea (2009), Perugia, Italy (2008), Kuala Lumpur, Malaysia (2007), Glasgow, UK (2006), Singapore (2005), Assisi, Italy (2004), Montreal, Canada (2003), and (as ICCS) Amsterdam, the Netherlands (2002) and San Francisco, USA (2001).

Computational Science constitutes the main pillar of most present research, industrial, and commercial applications, and plays a unique role in exploiting innovative ICT technologies, and the ICCSA conference series has, accordingly, provided ample opportunities to researchers and industry practitioners to discuss new ideas, to share complex problems and their solutions, and to shape new trends in Computational Science. As the conference mirrors society from a scientific point of view, this year's undoubtedly dominant theme was large language models, machine learning, and Artificial Intelligence,

and their applications across the most diverse technological, economic, and industrial fields, amongst others.

The ICCSA 2026 conference included 57 workshops, focusing on topical issues of utmost importance to science, technology, and society: from new computational approaches for earth science, to mathematical methods for image processing, new statistical and optimisation methods, several Artificial Intelligence approaches, sustainability issues, smart cities and related technologies, to name a few. In addition, ICCSA 2026 included six general tracks covering computational science and its applications: Computational Methods, Algorithms and Scientific Applications; High-Performance Computing and Networks; Geometric Modelling, Graphics and Visualisation; Advanced and Emerging Applications; Information Systems and Technologies; and Urban and Regional Planning.

For these Workshops Proceedings, we accepted 328 full papers, 23 short papers, and 3 Ph.D. Showcase papers from a total of 1184 submissions (acceptance rate 29.9%). Reviews were single blind, and submissions received 2.6 reviews each, on average. In the Main Conference Proceedings, we accepted 119 full papers, 3 short papers, and 2 Ph.D. Showcase papers from 431 submissions to the General Tracks of the Conference (acceptance rate of 29.9%). We would like to convey our sincere appreciation to the workshop chairs and co-chairs and program committee members, for their diligent work, commitment, and dedication.

The success and consistent maintenance of the ICCSA conference series in general, and of ICCSA 2026 in particular, rely upon the support of many people: authors, presenters, participants, keynote speakers, workshop chairs, session chairs, Organizing Committee members, student volunteers, Program Committee members, Advisory Committee members, International Liaison chairs, reviewers, and other individuals in various roles. Thus, we take this opportunity to wholeheartedly thank each and every one.

We also wish to thank the publisher, Springer, for accepting the proceedings, and for their kind assistance and cooperation during the editing process. We would cordially like to invite you to refer to the ICCSA website https://iccsa.org, where you can find the relevant details regarding this academic endeavour and our event.

June 2026

Chiara Garau
Ana Maria A. C. Rocha
Osvaldo Gervasi
Beniamino Murgante

Welcome Message from the Organizers

On behalf of the Local Organizing Committee of ICCSA 2026, it is our great pleasure to welcome you to the 26th International Conference on Computational Science and Its Applications, held from June 30 to July 3, 2026, at the University of Minho in Braga, Portugal.

We are deeply honoured by the trust placed in us by the ICCSA General Chairs in organizing another edition of this successful conference series. We hope that ICCSA 2026 provided an inspiring environment for scientific exchange, collaboration, and discussion among researchers from all over the world.

The conference took place at the University of Minho, one of the most prestigious higher education institutions in Portugal. Founded in 1973, the University of Minho has campuses in Braga and Guimaraes and is widely recognized for its academic excellence, research activity, and strong international outlook.

Braga, one of the oldest cities in Portugal, is renowned for its historical heritage, cultural vitality, and welcoming atmosphere.

Located in the heart of the Minho region, the city combines tradition with modernity, offering visitors renowned gastronomy, beautiful landscapes, and the unique character of *Vinho Verde* wine.

The scientific program of ICCSA 2026 included plenary lectures delivered by distinguished scientists, together with a broad range of workshops covering several areas of Computational Science and its Applications. We believe that the conference provided excellent opportunities to share ideas, discuss emerging challenges, and foster new collaborations.

In addition to the scientific sessions, participants had the opportunity to enjoy a rich social program.

Planned activities included a Welcome Reception at Largo do Paço, the Rectorate of the University of Minho, located in the former Archbishop's Palace - one of the most emblematic historical buildings in the city center; a *Vinho Verde* wine-tasting experience at Quinta d'Amares, where particiapnts could discover the art of winemaking; and a guided visit to the coastal city of Viana do Castelo, renowned for its maritime heritage, traditional architecture, and beautiful Atlantic scenery. The conference gala dinner took place at Colunata Eventos, located within the Bom Jesus do Monte sanctuary, a UNESCO World Heritage Site, where ICCSA participants enjoyed delicious recipes and a wonderful view over the city of Braga.

The success of ICCSA 2026 would not have been possible without the dedication and enthusiasm of many collaborators and volunteers, whose hard work and commitment greatly contributed to the organization of the conference.

On behalf of the Local Organizing Committee, it was our honour to warmly welcome you to the beautiful city of Braga and we hope you had an enjoyable and productive conference.

We enjoyed meeting you in Braga.

Sincerely yours,

On behalf of the Local Organizing Committee.

June 2026 Ana Maria A. C. Rocha

Organization

Honorary General Chairs

Bernady O. Apduhan	Kyushu Sangyo University, Japan
Kenneth C. J. Tan	Sardina Systems, UK

General Chairs

Ana Maria A. C. Rocha	University of Minho, Portugal
David Taniar	Monash University, Australia
Osvaldo Gervasi	University of Perugia, Italy

Program Committee Chairs

Beniamino Murgante	University of Basilicata, Italy
Chiara Garau	University of Cagliari, Italy

International Advisory Committee

Jemal Abawajy	Deakin University, Australia
Dharma P. Agarwal	University of Cincinnati, USA
Jagannath Aryal	University of Melbourne, Australia
Rajkumar Buyya	University of Melbourne, Australia
Claudia Bauzer Medeiros	University of Campinas, Brazil
Carl K. Chang	Iowa State University, USA
Elizabeth Chang	University of the Sunshine Coast, Australia
Manfred M. Fisher	Vienna University of Economics and Business, Austria
Pierre Frankhauser	University of Franche-Comté/CNRS, France
Marina L. Gavrilova	University of Calgary, Canada
Sumi Helal	University of Florida, USA & Lancaster University, UK
Bin Jiang	University of Gävle, Sweden
Yee Leung	Chinese University of Hong Kong, China
Biswajeet Pradhan	University of Technology Sydney, Australia
My Thai	University of Florida, USA

International Liaison Chairs

Giuseppe Borruso	University of Trieste, Italy
Elise De Donker	Western Michigan University, USA
Maria Noelia Faginas Lago	University of Perugia, Italy
Maria Irene Falcão	University of Minho, Portugal
Robert C. H. Hsu	Chung Hua University, Taiwan
Yeliz Karaca	UMass Chan Medical School, USA
Tae-Hoon Kim	Zhejiang University of Science and Technology, China
Vladimir Korkhov	Saint Petersburg University, Russia
Takashi Naka	Kyushu Sangyo University, Japan
Rafael D. C. Santos	National Institute for Space Research, Brazil
Maribel Yasmina Santos	University of Minho, Portugal
Anastasia Stratigea	National Technical University of Athens, Greece
Nguyen Canh Thai	Thuyloi University, Vietnam
Kh Md Nahiduzzaman	Mohammed VI Polytechnic University, Rabat, Morocco

Workshop and Session Organizing Chairs

Beniamino Murgante	University of Basilicata, Italy
Chiara Garau	University of Cagliari, Italy

Award Chair

Wenny Rahayu	La Trobe University, Australia

Publicity Committee Chairs

Elmer Dadios	De La Salle University, Philippines
Nataliia Kulabukhova	Saint Petersburg University, Russia
Daisuke Takahashi	Tsukuba University, Japan
Doan Quang Tu	Academy of Military Science and Technology, Vietnam
Shangwang Wang	Beijing University of Posts and Telecommunications, China

Local Organizing Committee Chairs

Maria Irene Falcão	University of Minho, Portugal
Ana Cristina Braga	University of Minho, Portugal
Rui José	University of Minho, Portugal
José Manuel Machado	University of Minho, Portugal
Jorge G. Rocha	University of Minho, Portugal

Technology Chair

Damiano Perri	University of Perugia, Italy

Workshops

Workshop on Advancements in Applied Machine-learning and Data Analytics (AAMDA 2026)

Workshop Organizers

Alessandro Costantini	INFN, Italy
Elisabetta Ronchieri	INFN CNAF, Italy

Workshop Program Committee Members

Luca Dell'Agnello	INFN, Italy
Barbara Martelli	INFN, Italy
Daniele Cesini	INFN CNAF, Italy
Giovanni Zurlo	INFN CNAF, Italy
Luca Giommi	INFN CNAF, Italy
Valentino Santucci	Università per Stranieri di Perugia, Italy
Francesco Santini	University of Perugia, Italy
Paolo Mengoni	Hong Kong Baptist University, China
Alessandro Bombini	Istituto Nazionale di Fisica Nucleare, Italy

Artificial Intelligence Tools In Managing Future Communities (AiMFutCOM 2026)

Workshop Organizers

Marco Vona	University of Basilicata, Italy
Francesco Scorza	University of Basilicata, Italy
Donatello Cardone	University of Basilicata, Italy
Maurizio Pollino	ENEA, Italy
Gregorio DAgostino	ENEA, Italy

Workshop Program Committee Members

Maria Danese	CNRISPC, Italy
Rachele Vanessa Gatto	Università della Basilicata, Italy
Simone Corrado	University of Basilicata, Italy
Alessandro Plaisant	University of Sassari, Italy
Sabrina Lai	University of Cagliari, Italy
Alfonso Annunziata	University of Basilicata, Italy
Jolanta Dvarioniene	Kaunas University of Technology, Lithuania

Advanced and Innovative Web Apps 2026 (AIWA 2026)

Workshop Organizers

Damiano Perri	University of Perugia, Italy
Osvaldo Gervasi	University of Perugia, Italy
Luciano Alessandro Ipsaro Palesi	University of Florence, Italy

Workshop Program Committee Members

David Berti	ART SpA, Italy
JungYoon Kim	Gachon University, South Korea
Taihoon Kim	Chonnam National University, South Korea

Advanced Mathematics and Computing Models in Complex Data-Intensive Computational Systems (AMCM 2026)

Workshop Organizers

Yeliz Karaca	UMass Chan Medical School and Massachusetts Institute of Technology, USA
Dumitru Baleanu	Lebanese American University, Lebanon
Osvaldo Gervasi	University of Perugia, Italy
Yudong Zhang	University of Leicester, UK
Majaz Moonis	UMass Chan Medical School, USA

Workshop Program Committee Members

Aziz Dursun	Virginia Tech University, USA
Akif Akgul	Hitit University, Turkey
Nizamettin Aydin	Istanbul Technical University, Turkey
Martin Bohner	Missouri University of Science and Technology, USA
Khan Muhammad	Sungkyunkwan University, South Korea

Advanced Numerical Approaches for assessment and design of no-tension Masonry Structures (ANAMS 2026)

Workshop Organizers

Antonino Iannuzzo	Universitá degli studi del Sannio, Italy
Natalia Pingaro	Politecnico di Milano, Italy
Andrea Montanino	International Centre for Numerical Methods in Engineering, Spain
Elham Mousavian	University of Edinburgh, UK
Stefano Galassi	University of Florence, Italy

Workshop Program Committee Members

Amal Gerges	University of Cagliari, Italy
Vittorio Paris	University of Bergamo, Italy
Alessandra Capolupo	Polytechnic University of Bari, Italy
Carlo Olivieri	Pegaso University, Italy
Generoso Vaiano	Pegaso University, Italy

Luigi Sibille	Norwegian University of Science and Technology, Norway
Francesco Picariello	Universitas Mercatorum, Italy
Ilaria Scarcelli	Politecnico di Bari, Italy
Fabian Bernal Orozco Barrera	National Autonomous University of Mexico, Mexico
Gabriel Barbat	Universitat Politècnica de Catalunya, Spain
Alejandro Cornejo	Universitat Politècnica de Catalunya, Spain
Mohammad Pourfouladi	Politecnico di Milano, Italy

Artificialization of land, Urban Transformation, and Territorial Resilience in a changing Climate (AUTTREC 2026)

Workshop Organizers

Celestina Fazia	Università di Enna Kore, Italy
Kh Md Nahiduzzaman	Mohammed VI Polytechnic University, Morocco
Hassan Radoine	Mohammed VI Polytechnic University, Morocco
Cezar Morar	University of Oradea, Romania
Antonio García Martínez	Universidad de Sevilla, Spain
Clara Stella Vicari Aversa	Università degli Studi Mediterranea di Reggio Calabria, Italy

Workshop Program Committee Members

Federica Sortino	Universitià degli studi di Enna Kore, Italy
Giulia Fernanda Grazia Catania	Universit di Enna Kore, Italy
Maria Francesca Faro	Ministry of Education and Merit (MIM), Italy
Aleksandar Valjarević	University of Belgrade, Serbia
Laila El Ghazouani	Mohammed VI Polytechnic University, Morocco
Adnane Labbaci	Ibn Tofail University, Morocco

Blockchain and Distributed Ledgers: Technologies and Applications (BDLTA 2026)

Workshop Organizers

Vladimir Korkhov	Saint Petersburg State University, Russia
Elena Stankova	Saint Petersburg State University, Russia
Nataliia Kulabukhova	AutoTech, Russia

Workshop Program Committee Members

Vasily Solodkov	National Research University Higher School of Economics, Moscow, Russia
Alexander Bogdanov	St. Petersburg State University, Russia
Alexander Degtyarev	Saint Petersburg State University, Russia
Dmitrii Vasiunin	Deutsche Telekom Cloud Services E.P.E., Greece
Serob Balyan	Institute for Informatics and Automation Problems, National Academy of Sciences of Armenia, Armenia
Suren Abrahamyan	Institute for Informatics and Automation Problems of NAS RA, Armenia
Ashot Gevorkyan	Institute for Informatics and Automation Problems NAS of Armenia, Armenia
Martin Vala	Pavol Jozef Šafárik University, Slovak Republic
Nodir Zaynalov	Samarkand Branch of Tashkent University of Information Technologies Named After Muhammad al Khwarizmi, Uzbekistan
Michail Panteleyev	Saint Petersburg Electrotechnical University, Russia
Alexander Degtyarev	St. Petersburg State University, Russia
Alexander Bogdanov	St. Petersburg State University, Russia

Bio and Neuro inspired Computing and Applications (BIONCA 2026)

Workshop Organizers

Nadia Nedjah	State University of Rio de Janeiro, Brazil
Luiza de Macedo Mourelle	State University de Rio de Janeiro, Brazil

Workshop Program Committee Members

Alan Oliveira	Universidade de Lisboa, Portugal
Rogerio Calazan	Instituto de Estudos do Mar Almirante Paulo Moreira - Brazilian Navy, Brazil
Marcos Farias	Instituto de Engenharia Nuclear, Brazil
Joelmir Ramos	State University de Rio de Janeiro, Brazil
Luigi Maciel Ribeiro	State University of Rio de Janeiro, Brazil

Computational and Applied Mathematics (CAM 2026)

Workshop Organizers

Maria Irene Falcao	University of Minho, Portugal
Fernando Miranda	Universidade do Minho, Portugal

Workshop Program Committee Members

Graca Tomaz	Instituto Politécnico da Guarda, Portugal
Helmuth Malonek	University of Aveiro, Portugal
Isabel Cacao	University of Aveiro, Portugal
Joao Morais	Autonomous Technological Institute of Mexico (ITAM), Mexico
Lidia Aceto	Università del Piemonte Orientale, Italy
Luís Ferrás	University of Minho, Portugal
Patricia Beites	University of Beira Interior, Portugal
Paulo Amorim	Escola de Matemática Aplicada da Fundação Getulio Vargas, Brazil
Regina de Almeida	University of Trás-os-Montes and Alto Douro, Portugal
Ricardo Severino	University of Minho, Portugal

Computational and Applied Statistics for Data & AI (CAS 2026)

Workshop Organizers

Ana Cristina Braga	University of Minho, Portugal
Ines Sena	University of Minho, Portugal
Joao Mendes	Polytechnic Institute of Bragança, Portugal

Workshop Program Committee Members

A. Manuela Goncalves	University of Minho, Portugal
Vera Afreixo	University of Aveiro, Portugal
Ivan Rodriguez Conde	University of Arkansas at Little Rock, USA
Isabel Natário	Universidade Nova de Lisboa, Portugal
Luca Parodi	University of Genoa, Italy
Flávia Pires	Instituto Politecnico de Bragança, Portugal
Beatriz Flamia Azevedo	Instituto Politecnico de Bragança, Portugal
Adriano Silva	Instituto Politecnico de Bragança VAT 600013758, Portugal

Ana Paula Amorim	Universidade do Minho, Portugal
Angela Silva	Universidade do Minho, Portugal
Arunasalam Rahunanthan	Central State University, USA
Bruna Ramos	University of Minho, Portugal
Cecília Maria Vasconcelos Costa e Castro	Universidade do Minho, Portugal
Florbela Correia	Polytechnic Institute of Viana do Castelo, Portugal
Joaquim Goncalves	Polytechnic Institute of Cávado and Ave, Portugal
Lino Costa	University of Minho, Portugal
Luís Miguel Matos	Universidade do Minho, Portugal
Marco Reis	University of Coimbra, Portugal
Maria Mourao	Polytechnic Institute of Viana do Castelo, Portugal
Martin Perez Perez	University of Vigo, Spain
Sandra Aleixo	FCiências.ID, Portugal
Paula Odete Fernandes	Instituto Politecnico de Bragança, Portugal
Iván Rodríguez Conde	Universidade de Vigo, Spain
Telmo Miguel Pires Pinto	University of Minho, Portugal

Circular Economy and Territorial Competitiveness (CETC 2026)

Workshop Organizers

Gerardo Carpentieri	University of Naples Federico II, Italy
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Floriana Zucaro	University of Naples Federico II, Italy
Zhadyra Konurbaeva	D. Serikbayev East Kazakhstan Technical University, Kazakhstan
Saule Kumargazhanova	D. Serikbayev East Kazakhstan Technical University, Kazakhstan

Workshop Program Committee Members

Rosaria Battarra	ISMed-CNR, Italy
Laura Ascione	Università degli studi di Napoli Federico II, Italy
Valerio Martinelli	Università di Napoli Federico II, Italy
Madina Yussubaliyeva	D. Serikbayeva East Kazakhstan Technical University, Kazakhstan
Samat Baigereyev	D. Serikbayev East Kazakhstan Technical University, Kazakhstan

Circular Ecological Heritage Communities: Regeneration and Evaluation Strategies for Historic Villages, Landscapes and Rural Territories (CHEC 2026)

Workshop Organizers

Mariarosaria Angrisano	Pegaso University, Italy
Antonia Gravagnuolo	CNR Institute of Heritage Science, Italy
Francesca Buglione	CNR ISPC, Italy
Piera Della Morte	Pegaso Telematic University, Italy

Workshop Program Committee Members

Giulia Datola	Politecnico di Milano, Italy
Marco Rossitti	Politecnico di Milano, Italy
Federico DellAnna	Politecnico di Torino, Italy
Grazia Neglia	Pegaso Telematic University, Italy
Vanessa Assumma	University of Bologna, Italy
Ippolita Mecca	Pegaso Telematic University, Italy
Ferdinando Verardi	Pegaso Telematic University, Italy
Eleonora Sottile	University of Cagliari, Italy
Dina Jovanovic	Politecnico di Milano, Italy
Martina Mulinacci	Politecnico di Milano, Italy
Davide Cerati	Politecnico di Milano, Italy
Daniela Oreni	Politecnico di Milano, Italy
Patrizia Riganti	University of Glasgow, UK
Carlo Olivieri	Pegaso University, Italy
Giulio Cavana	Politecnico di Torino, Italy
Maria Tartari	CNR, Italy
Luigi Fusco Girard	Pegaso Telematic University, Italy
Albina Scioti	Università Telematica Pegaso, Italy
Eleonora Righetto	Università degli Studi di Padova, Italy
Chiara D'Alpaos	University of Padua, Italy
Anthony Suppa	Politecnico di Torino, Italy
Giorgia Malavasi	Politecnico di Torino, Italy
Kaoutare Amini Alaoui	Mohamed VI Polytechnic University, Morocco
Elena Todella	University of Parma, Italy
Leonhard Seyboth	Federal University of Santa Catarina, Brazil
Jianzhuo Xu	Politecnico di Milano, Italy
Ludovica Nasca	University of Catania, Italy
Alessandro Plaisant	University of Sassari, Italy

Maria Carla Saliu	University of Cagliari, Italy
Ons Zoghlami	Sapienza University of Rome, Italy
Irena Ištoka Otković	Josip Juraj Strossmayer University of Osijek, Croatia
Daniele Cannatella	TU Delft, Italy
Tanja Congiu	University of Sassari, Italy
Juan Camilo Giraldo	Institute for Advanced Architecture of Catalonia, Spain
YiChang Liu	Bio-Architecture Formosana, Taiwan
Yasmine Selim	Ain Shams University, Egypt
Simona Tondelli	Alma Mater Studiorum University of Bologna, Italy
Efrossyni Tsakiri	National Technical University of Athens, Greece

Cyber Intelligence and Applications (CIA 2026)

Workshop Organizers

Gianni D'Angelo	University of Salerno, Italy
Francesco Palmieri	University of Salerno, Italy
Massimo Ficco	University of Salerno, Italy

Workshop Program Committee Members

Arcangelo Castiglione	University of Salerno, Italy
Ugo Fiore	University of Salerno, Italy
Michele Mastroianni	University of Foggia, Italy
Marcello Trovati	Edge Hill University, UK
Raffaele Esposito	University of Salerno, Italy
Marek Ogiela	AGH University of Science and Technology, Poland
Chang Choi	Gachon University, South, South Korea
Mauro Migliardi	University of Padua, Italy

Computational Methods, Statistics and Industrial Mathematics (CMSIM 2026)

Workshop Organizers

Maria Filomena Teodoro	CINAV, Escola Naval, Instituto Universitário Militar, Portugal

Paula Simões	Academia Militar, Portugal
Teresa A. Oliveira	University of Lisbon and Universidade Aberta, Portugal

Workshop Program Committee Members

Luis Teodoro	University of Oslo, Norway
Eliana Costa e Silva	Polytechnic of Porto, Portugal
Ana Borges	Instituto Politécnico do Porto, Portugal
Cristina Lopes	Polytechnic Institute of Porto, Portugal
Fernando Lucas Carapau	University of Évora, Portugal
Antonio S. Macias	University of A Coruña, Spain
Matteo Sacchet	University of Turin, Italy
Marina Marchisio Conte	University of Turin, Italy
Christos Kitsos	University of West Attica, Greece
M. Rosário Ramos	Universidade Aberta, University of Lisbon, Portugal
António Pacheco	Instituto Superior Técnico, Portugal
Ana Sofia Rézio	Polytechnic Institute of Lisbon, Portugal
Paula Simões	Military Academy Portuguese Army, Portugal
Ricardo Moura	NOVA University of Lisbon, Portugal
Ana Oliveira	Academia Militar, Portugal
Sílvia Chá	Military Academy, Portugal
Carlos Rodrigues	Instituto Politecnico de Bragança, Portugal

Computational Optimization and Applications (COA 2026)

Workshop Organizers

Ana Maria A. C. Rocha	University of Minho, Portugal
Humberto Rocha	University of Coimbra, Portugal

Workshop Program Committee Members

Eligius Hendrix	Universidad de Málaga, Spain
Joana Dias	University of Coimbra, Portugal
Ana Pereira	Polytechnic Institute of Bragança, Portugal
Lino Costa	University of Minho, Portugal, Portugal
Filipe Alvelos	Universidade do Minho, Portugal
Ana Paula Teixeira	Universidade do Minho, Portugal

Florbela P. Fernandes	Politecnico de Bragança, Portugal
Ana Monteiro	University of Coimbra, Portugal
Rui Pascoal	University of Coimbra, Portugal
Rommel Regis	Saint Joseph's University, USA
Telmo Miguel Pires Pinto	University of Minho, Portugal
Emerson Paiva	Minho University, Portugal

Computational Methods in Advanced Networking Technologies (CoMANT 2026)

Workshop Organizers

Eslam Farsimadan	University of Salerno, Italy
Leila Moradi	University of Salerno, Italy
Francesco Palmieri	University of Salerno, Italy

Workshop Program Committee Members

Massimo Ficco	University of Salerno, Italy
Ugo Fiore	University of Salerno, Italy
Michele Mastroianni	University of Foggia, Italy
Arcangelo Castiglione	University of Salerno, Italy
Francesco Giuseppe Zammarrelli	University of Salerno, Italy
Annalaura Miglino	University of Salerno, Italy
Mojtaba Sharifzadeh	Università di Trento, Italy
Zahra Ebadi Ansaroudi	Fondazione Bruno Kessler, Italy
Mohammad Vazir Panah	University of Salerno, Italy
Gennaro Pio Rimoli	Università degli Studi di Salerno, Italy

Computational Astrochemistry 2026 (CompAstro 2026)

Workshop Organizers

Marzio Rosi	University of Perugia, Italy
Daniela Ascenzi	Università degli Studi di Trento, Italy
Nadia Balucani	University of Perugia, Italy
Stefano Falcinelli	University of Perugia, Italy

Workshop Program Committee Members

Luca Mancini	University of Perugia, Italy
Andrea Giustini	University of Perugia, Italy
Albert Rimola	Universitat Autònoma de Barcelona, Spain
Piero Ugliengo	Università degli Studi di Torino, Italy
Dario Campisi	University of Perugia, Italy
Gianmarco Valuzzo	University of Perugia, Italy
Giacomo Pannacci	Università degli Studi di Perugia, Italy
Giacomo Giorgi	Università degli Studi di Perugia, Italy
Gabriella Di Genova	Università degli Studi di Perugia, Italy
Lisa Giani	Université Grenoble Alpes, IPAG, France
Emilia Valenca Ferreira de Aragao	University of Montpellier, CEA, CNRS, ENSCM, France
Marco Parriani	Università degli studi di Perugia, Italy
Costanza Borghesi	CNR Istituto Nanoscienze, Italy

Computational methods for Porous materials (CompPor 2026)

Workshop Organizers

Vadim Lisitsa	Institute of Petroleum Geology and Geophysics SB RAS, Russia
Evgeniy Romenski	Sobolev Institute of Mathematics, Russia

Workshop Program Committee Members

Tatyana Khachkova	Institute of Petroleum Geology and Geophysics SB RAS, Russia
Dmitry Prokhorov	Institute of Petroleum Geology and Geophysics SB RAS, Russia
Mikhail Novikov	Sobolev Institute of Mathematics SB RAS, Russia
Galina Reshetova	Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
Sergey Solovyev	Sobolev Institute of Mathematics SB RAS, Russia
Vladimir Cheverda	Sobolev Institute of Mathematics SB RAS, Russia

Workshop on Computational Science and HPC (CSHPC 2026)

Workshop Organizers

Elise de Doncker	Western Michigan University, USA
Hideo Matsufuru	High Energy Accelerator Research Organization, Japan
Robert Makin	Western Michigan University, USA

Workshop Program Committee Members

Hiroshi Daisaka	Hitotsubashi University, Japan
James Rhodes	Western Michigan University, USA
Naohito Nakasato	University of Aizu, Japan
Issaku Kanamori	RIKEN, Japan
Takahiro Ueda	Juntendo University, Japan
Keigo Nitadori	RIKEN, Japan
Fukuko Yuasa	KEK, Japan
Shigeyoshi Aoyama	KEK, Japan
Tatsumi Aoyama	University of Tokyo, Japan
Tasnim Gharaibeh	Kalamazoo College, USA

Cities, Technologies and Planning (CTP 2026)

Workshop Organizers

Giuseppe Borruso	University of Trieste, Italy
Beniamino Murgante	University of Basilicata, Italy
Malgorzata Hanzl	Lodz University of Technology, Poland
Anastasia Stratigea	National Technical University of Athens, Greece
Ljiljana Zivkovic	Republic Geodetic Authority, Serbia
Ginevra Balletto	University of Cagliari, Italy

Workshop Program Committee Members

Silvia Battino	University of Sassari, Italy
Mara Ladu	University of Cagliari, Italy
Maria del Mar Munoz	University of Cádiz, Spain
Ainhoa Amaro	EVOCONS, Spain
Enrico Dagostini	University of Malta, Malta

Francesca Krasna	Università degli Studi di Trieste, Italy
Brisol García García	Universidad Politécnica de Quintana Roo, Mexico
Giovanni Mauro	University of Campania Luigi Vanvitelli, Italy
Maria Ronza	University of Naples Federico II, Italy
Clara Di Fazio	University of Trieste, Italy
Andrea Gallo	Università degli Studi di Trieste, Italy
Francesca Sinatra	University of Trieste, Italy

Evaluating Inner Areas Potentials (EIAP 2026)

Workshop Organizers

Diana Rolando	Politecnico di Torino, Italy
Manuela Rebaudengo	Politecnico di Torino, Italy
Alice Barreca	Politecnico di Torino, Italy
Giorgia Malavasi	Politecnico di Torino, Italy

Workshop Program Committee Members

John Accordino	Virginia Commonwealth University, USA
Sara Favargiotti	University of Trento, Italy
Barbara Lino	Università degli Studi di Palermo, Italy
Maddalena Ferretti	Università Politecnica delle Marche, Italy
Eugenio Muccio	University of Naples Federico II, Italy
Giuliano Poli	University of Naples Federico II, Italy
Giulia Datola	Politecnico di Milano, Italy
Elena Todella	University of Parma, Italy
Asja Aulisio	Politecnico di Torino, Italy
Marco Rossitti	Politecnico di Milano, Italy
Martina Spinelli	Politecnico Torino, Italy
Amina Pereno	Politecnico di Torino, Italy
Francesco Bruzzone	Università Iuav di Venezia, Italy
Alexandra Stankulova	Politecnico di Torino, Italy
Benedetta Di Leo	Università Politecnica delle Marche, Italy
Edda Donati	University of Rome Tor Vergata, Italy
Vanessa Assumma	University of Bologna, Italy
Beatrice Mecca	Politecnico di Torino, Italy

sustainablE mobiLity and Last mIle lOgisTic in urban context (ELLIOT 2026)

Workshop Organizers

Tiziana Campisi	University of Enna Kore, Italy
Socrates Basbas	Aristotle University of Thessaloniki, Greece
Florin Rusca	National University of Science and Technology Politehnica Bucharest, Romania
Grigorios Fountas	Aristotle University of Thessaloniki, Greece
Aleksandra Deluka-Tibljas	University of Rijeka, Croatia

Workshop Program Committee Members

Francesco Russo	University of Reggio Calabria, Italy
Oana Dinu	National University of Science and Technology Politehnica Bucharest, Romania
Antonio Polimeni	Università di Messina, Italy
Antonio Russo	University of Enna Kore, Italy
Elena Cocuzza	Univesity of Catania, Italy
Corrado Rindone	Università degli studi Mediterranea di Reggio Calabria, Italy

Econometrics and Multidimensional Evaluation in Urban Environment (EMEUE 2026)

Workshop Organizers

Maria Cerreta	University of Naples Federico II, Italy
Giuseppe Ciciriello	University of Naples Federico II, Italy
Simona Panaro	University of Sussex Business School, UK
Piero Zizzania	Università of Naples Federico II, Italy
Pierluigi Morano	Polytechnic University of Bari, Italy
Francesco Tajani	Sapienza University of Rome, Italy

Workshop Program Committee Members

Giuliano Poli	Università Federico II di Napoli, Italy
Ludovica La Rocca	University of Naples Federico II, Italy
Sabrina Sacco	University of Naples Federico II, Italy

Ecosystem Services in Spacial Planning for climate neutral urban and rural areas (ESSP 2026)

Workshop Organizers

Sabrina Lai	University of Cagliari, Italy
Carmela Gargiulo	Università Federico II di Napoli, Italy
Beniamino Murgante	University of Basilicata, Italy
Francesco Scorza	University of Basilicata, Italy
Corrado Zoppi	University of Cagliari, Italy
Floriana Zucaro	University of Naples Federico II, Italy

Workshop Program Committee Members

Alfonso Annunziata	University of Basilicata, Italy
Ginevra Balletto	University of Cagliari, Italy
Ivan Blečić	University of Cagliari, Italy
Giuseppe Borruso	University of Trieste, Italy
Barbara Caselli	Università degli Studi di Parma, Italy
Maria Cerreta	University of Naples Federico II, Italy
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Federica Isola	University of Cagliari, Italy
Francesca Leccis	University of Cagliari, Italy
Federica Leone	University of Cagliari, Italy
Silvia Rossetti	Università degli Studi di Parma, Italy
Luigi Santopietro	Institute of Methodologies for Environmental Analysis CNR IMAA, Italy
Simone Corrado	University of Basilicata, Italy
Rachele Vanessa Gatto	Università della Basilicata, Italy

The 16th International Workshop on Future Information System Technologies and Applications (FiSTA 2026)

Workshop Organizers

Bernady O. Apduhan	Kyushu Sangyo University, Japan
Rafael Santos	National Institute for Space Research (INPE), Brazil

Workshop Program Committee Members

Agustinus Waluyo	La Trobe University, Australia
Earl Ryan Aleluya	Mindanao State University Iligan Institute of Technology, Philippines
Eric Pardede	La Trobe University, Australia
Kai Cheng	Kyushu Sangyo University, Japan
Fenghui Yao	Tennessee State University, USA
Yusuke Gotoh	Okayama University, Japan
Kazuaki Tanaka	Kyushu Institute of Technology, Japan
Toshihiro Yamauchi	Okayama University, Japan
Toshihiro Uchibayashi	Kyushu University, Japan
Krishnamoorthy Ranganathan	Chennai Institute of Technology, India
Orven Llantos	MSU-Iligan Institute of Technology, Philippines
Anton Louise De Ocampo	Batangas State University, Philippines

Geographical Analysis, Urban Modeling, Spatial Statistics (Geog-And-Mod 2026)

Workshop Organizers

Giuseppe Borruso	University of Trieste, Italy
Beniamino Murgante	University of Basilicata, Italy
Hartmut Asche	Hasso Plattner Institute for Digital Engineering, Germany
Rodrigo Tapia-McClung	Centro de Investigación en Ciencias de Información Geoespacial, Mexico
Andreas Fricke	Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany

Workshop Program Committee Members

Ginevra Balletto	University of Cagliari, Italy
Mara Ladu	University of Cagliari, Italy
Gabriella Schoier	Università di Trieste, Italy
Marco Mazzarino	Iuav University of Venice, Italy
Malgorzata Hanzl	Lodz University of Technology, Poland
Anastasia Stratigea	National Technical University of Athens, Greece
Maria del Mar Munoz	University of Cádiz, Spain
Veronica Camerada	Università degli Studi di Sassari, Italy

Generative and Neuro-Symbolic Artificial Intelligence for Real-World Systems (GNS-AI 2026)

Workshop Organizers

Hugo Peixoto	University of Minho, Portugal
Manuel Rodrigues	University of Minho, Portugal
Regina Sousa	University of Minho, Portugal

Workshop Program Committee Members

Dalila Duraes	Universidade do Minho, Portugal
Paulo Novais	Universidade do Minho, Portugal
Laura Cavalcante	Instituto Superior de Engenharia do Porto, Portugal
Diogo Martinho	Instituto Superior de Engenharia de Porto, Portugal
Hélia Guerra	University of the Azores, Portugal
Rohit Raja	Guru Ghasidas Vishwavidyalaya, India
Hien Nguyen	University of Information Technology, Vietnam
Antonio Carlos Da Silva Abelha	Universidade do Minho, Portugal
Cristiana Neto	University of Minho, Portugal
José Machado	University of Minho, Portugal
Luís Conceicao	Instituto Superior de Engenharia do Porto, Portugal
Omar Gomez	Universidad Nacional de Colombia, Colombia
Luis Mendes Gomes	University of the Azores, Portugal
Nicolas Lori	University of the Azores, Portugal
Sara Balderas Díaz	University of Cádiz, Spain
Gabriel Guerrero Contreras	University of Cádiz, Spain
Francisco Zayas Gato	Universidade da Coruña, Spain
Diana Ferreira	University of Minho, Portugal
Jolanta Miliauskaite	Vilnius University, Lithuania

Geomatics for Resource Monitoring and Management (GRMM 2026)

Workshop Organizers

Alberico Sonnessa	Politecnico di Bari, Italy
Eufemia Tarantino	Politecnico di Bari, Italy
Marco Locurcio	Polytechnic University of Bari, Italy

Workshop Program Committee Members

Valeria Belloni	Sapienza University of Rome, Italy
Roberta Ravanelli	University of Liège, Belgium
Michela Ravanelli	Sapienza Università di Roma, Italy
Alessandra Mascitelli	University of Chieti Pescara, Italy
Francesco Bimbo	University of Foggia, Italy
Francesco Di Capua	University of Basilicata, Italy
Gaetano Falcone	Università degli Studi di Napoli Federico II, Italy
Francesco Tajani	Sapienza University of Rome, Italy
Paola Amoruso	LUM Giuseppe Degennaro University, Italy
Pierfrancesco De Paola	University of Naples Federico II, Italy
Isidora Đurić	University of Novi Sad, Serbia
Rossella Petti	Politecnico di Bari, Italy
Valeria Cera	University of Naples Federico II, Italy
Antonino Iannuzzo	Universitá degli studi del Sannio, Italy
Pier Francesco Recchi	Università di Napoli Federico II, Italy
Giuseppe Francesco Cesare Lama	University of Naples Federico II, Italy
Ester Scotto di Perta	University of Naples Federico II, Italy
Antonino Maltese	University of Palermo, Italy
Andrea Montanino	University of Naples, Italy
Carla Cavallo	University of Salerno, Italy
Dario Esposito	Polytechnic University of Bari, Italy
Rosaria Parente	Universitas Mercatorum, Italy
Claudia Panciera	Water Research Institute (IRSA) CNR, Italy
Pasquale Perrini	Consorzio Interuniversitario per l'Idrologia, Italy
Beniamino Onorati	Università degli Studi della Basilicata, Italy
Francesco Sica	Sapienza University of Rome, Italy

International Workshop on Information and Knowledge in the Internet of Things (IKIT 2026)

Workshop Organizer

Teresa Guarda	Universidad Estatal Península de Santa Elena, Ecuador

Workshop Program Committee Members

Isabel Lopes	Instituto Politécnico de Bragança, Portugal
Luis Mazon	Bitrum Research Group, Spain

Arnulfo Alanis	Tecnológico Nacional de México, Campus Tijuana, Mexico
Filipe Mota Pinto	Instituto Politécnico de Leiria, Portugal
Luis Enrique Chuquimarca Jiménez	Universidad Estatal Penìnsula de Santa Elena, Ecuador
Gustavo Gatica	Universidad Andres Bello, Chile
Jairo Coronado Hernandez	Universidad de la Costa, Colombia
Maria Fernanda Augusto	BITrum Research Center, Spain
José María Díaz Nafría	Universidad a Distancia de Madrid, Spain
Modestos Stavrakis	University of the Aegean, Greece
Isabel Ribeiro	Instituto Politécnico Bragança, Portugal
Bruno Sousa	University of Coimbra, Portugal
Datzania Villao	Universidad Estatal Península de Santa Elena, Ecuador

Intelligent Sustainable Development: from Models to Real-World Impact (ISD 2026)

Workshop Organizers

Diego Brandão	Federal Center for Technological Education of Rio de Janeiro, Brazil
Angela Ferreira	Polytechnic University of Bragança, Portugal
Eduardo Bezerra	Federal Center for Technological Education of Rio de Janeiro, Brazil
Raphael Guerra	Universidade Federal Fluminense, Brazil
Douglas Cardoso	University of Coimbra, Portugal
Rui Pedro Lopes	Instituto Politecnico de Bragança, Portugal

Workshop Program Committee Members

Wagner Telles	Universidade Federal Fluminense, Brazil
Jader Lugon Junior	Instituto Federal Fluminense, Brazil
Edcarllos Santos	Universidade Federal do Sul da Bahia, Brazil
Kennedy Fernandes	Universidade Federal de Juiz de Fora, Brazil
Israel Mendonca	Kumamoto University, Japan
Pedro Henrique Gonzalez	Federal University of Rio de Janeiro, Brazil
Pablo Luiz Araujo Munhoz	Federal University of Ouro Preto, Brazil
Elloa Guedes	Amazonas State University, Brazil
Danielli Araújo Lima	Instituto Federal do Triângulo Mineiro, Brazil
Ary Henrique Oliveira	Universidade Federal de Tocantins, Brazil

Paula Odete Fernandes	Instituto Politecnico de Bragança, Portugal
Hiago Rocha	National Laboratory for Scientific Computing, Brazil
Andre Luiz Coutinho Merlo	Centro Federal de Educacao Tecnologica Celso Suckow da Fonseca CEFET RJ, Brazil
Kennedy Fernandes	Universidade Federal de Juiz de Fora, Brazil
Ary Henrique Oliveira	Universidade Federal de Tocantins, Brazil
Mayara Omai	Centro Federal de Educacao Tecnologica Celso Suckow da Fonseca CEFET RJ, Brazil

Multidimensional Evolutionary Evaluations for Transformative Approaches (MEETA 2026)

Workshop Organizers

Maria Cerreta	University of Naples Federico II, Italy
Giuliano Poli	Università Federico II di Napoli, Italy
Gaia Daldanise	CNR Institute for Research on Innovation and Services for Development, Italy
Ludovica La Rocca	University of Federico II of Naples, Italy
Maria Somma	University of Naples Federico II, Italy
Sabrina Sacco	University of Naples Federico II, Italy

Workshop Program Committee Members

Chiara Mazzarella	TUDelft, The Netherlands
Simona Panaro	University of Sussex Business School, UK
Eugenio Muccio	University of Naples Federico II, Italy

Building multi-dimensional Models for assessing complex Environmental Systems (MES 2026)

Workshop Organizers

Vanessa Assumma	University of Bologna, Italy
Caterina Caprioli	Politecnico di Torino, Italy
Giulia Datola	Politecnico di Milano, Italy
Federico DellAnna	Politecnico di Torino, Italy
Marta Dell'Ovo	Politecnico di Milano, Italy
Marco Rossitti	Politecnico di Milano, Italy

Workshop Program Committee Members

Mariarosaria Angrisano	Pegaso University, Italy
Giuliano Poli	University of Naples Federico II, Italy
Francesco Sica	Sapienza University of Rome, Italy
Eugenio Muccio	University of Naples Federico II, Italy
Sebastiano Barbieri	Politecnico di Torino, Italy
Giulio Cavana	Politecnico di Torino, Italy
Marta Bottero	Politecnico di Torino, Italy
Francesca Torrieri	Politecnico di Milano, Italy
Maksims Feofilovs	Riga Technical University, Latvia
Daniela Maiullari	Delft University of Technology, The Netherlands
Elena Todella	University of Parma, Italy
Yvonne Russo	University of Campania L. Vanvitelli, Italy
Tommaso Lazzarin	University of Padua, Italy
Francesco Tajani	Sapienza University of Rome, Italy
Claudia Amadei	University of Padua, Italy
Isabella Maria Lami	Politecnico di Torino, Italy
Federico Vagnon	Politecnico di Torino, Italy
Francesca Torrieri	Politecnico di Milano, Italy
Elena De Angelis	Politecnico di Torino, Italy
Piergiorgio Farfariello	Politecnico di Torino, Italy
Francesca Biondi	Politecnico di Torino, Italy
Giulio Cavana	Politecnico di Torino, Italy
Fabiana Forte	Università della Campania Luigi Vanvitelli, Italy
Siyu Qiu	University of Bologna, Italy
Eugenio Muccio	University of Naples Federico II, Italy
Iuliia Kozlova	University of Bologna, Italy
Danny Casprini	Politecnico di Milano, Italy
Franco Corti	University of Padua, Italy
Elisa Sbardella	Politecnico di Torino, Italy
Andrea Sarcina	Università di Padova, Italy
Daniela Menna	Università degli studi della Campania Luigi Vanvitelli, Italy
Sebastiano Barbieri	Politecnico di Torino, Italy
Ana Soares	University of Minho, Portugal
Francesco Sica	Sapienza University of Rome, Italy
Chiara D'Alpaos	University of Padua, Italy
Alessandra Oppio	Politecnico di Milano, Italy
Maria Rosaria Guarini	Sapienza University of Rome, Italy
Alessia Crisopulli	Politecnico di Milano, Italy
Elisabetta Colucci	Politecnico di Torino, Italy

Giorgia Sugoni	Links Foundation, Italy
Lucia Petronio	University of Bologna, Italy
Marco Alvise Bragadin	University of Bologna, Italy

Computational Mechanics, Intelligence, and Data (MIND 2026)

Workshop Organizers

Lelio Campanile	Università degli Studi della Campania Luigi Vanvitelli, Italy
Renato Zona	Università della Campania Vanvitelli, Italy

Workshop Program Committee Members

Vincenzo Minutolo	Università degli Studi della Campania Luigi Vanvitelli, Italy
Fiammetta Marulli	University of Campania Luigi Vanvitelli, Italy
Simone Palladino	Università degli Studi della Campania Luigi Vanvitelli, Italy
Ali Forooghi	University of Campania Luigi Vanvitelli, Italy
Mario Argenziano	University of Palermo, Italy
Donato Perfetto	University of Campania Luigi Vanvitelli, Italy

Modelling Liveable Cities: Techniques, Methods, Challenges, and Perspectives Behind the X-Minute City (MLC 2026)

Workshop Organizers

Federico Mara	University of Pisa, Italy
Valerio Cutini	University of Pisa, Italy

Workshop Program Committee Members

Beniamino Murgante	University of Basilicata, Italy
Chiara Garau	University of Cagliari, Italy
Francesco Zullo	University of LAquila, Italy
Alessandro Marucci	University of L Aquila, Italy
Chiara Di Dato	University of L Aquila, Italy
Giampiero Lombardini	University of Genoa, Italy

Alfonso Annunziata	University of Basilicata, Italy
Simone Corrado	University of Basilicata, Italy
Müslüm Hacar	Yildiz Technical University, Turkey
Federica Deri	University of Pisa, Italy
Chiara Anselmi	University of Pisa, Italy

Mathematical Methods for Image Processing and Understanding 2026 (MMIPU 2026)

Workshop Organizers

Ivan Gerace	Università degli Studi di Perugia, Italy
Gianluca Vinti	Università degli Studi di Perugia, Italy
Arianna Travaglini	Lucian Blaga University of Sibiu, Romania

Workshop Program Committee Members

Marco Baioletti	University of Perugia, Italy
Marco Donatelli	University of Insubria, Italy
Muhammad Hanif	Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan
Francesco Marchetti	University of Padua, Italy
Wolfgang Erb	Università degli Studi di Padua, Italy
Ana Maria Acu	Lucian Blaga University of Sibiu, Romania
Simone Rebegoldi	University of Modena and Reggio Emilia, Italy
Valentina Giorgetti	Università degli Studi di Perugia, Italy
Vita Leonessa	University of Basilicata, Italy

MOdels and indicators for assessing and measuring the urban settlement deVElopment in the view of NET ZERO by 2050 (MOVEto0 2026)

Workshop Organizers

Lorena Fiorini	University of L'Aquila, Italy
Lucia Saganeiti	Institute of Methodologies for Environmental Analysis (IMAA) CNR, Italy
Angela Pilogallo	Institute of Methodologies for Environmental Analysis (IMAA) CNR, Italy

Francesco Zullo	University of L'Aquila, Italy
Alessandro Marucci	University of L'Aquila, Italy

Workshop Program Committee Members

Ginevra Balletto	University of Cagliari, Italy
Giuseppe Borruso	University of Trieste, Italy
Chiara Garau	University of Cagliari, Italy
Beniamino Murgante	University of Basilicata, Italy
Ljiljana Zivkovic	Republic Geodetic Authority, Serbia
Luigi Santopietro	Institute of Methodologies for Environmental Analysis National Research Council of Italy CNR IMAA, Italy
Ilaria Delponte	University of Genoa, Italy
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Chiara Di Dato	University of L'Aquila, Italy

Numerical Aspects in Machine Learning For Dynamical Systems (NAMLYS 2026)

Workshop Organizers

Pasquale De Luca	Parthenope University of Naples, Italy
Livia Marcellino	University of Naples Parthenope, Italy
Valeria Mele	University of Naples Federico II, Italy

Workshop Program Committee Members

Dario Caramiello	University of Naples Parthenope, Italy
Giuseppe Fiorillo	Parthenope University of Naples, Italy
Marco Lettiero	University of Naples Parthenope, Italy
Francesco Peluso	University of Naples Parthenope, Italy

Numerical Methods and Machine Learning in Science & Engineering (NMLSE 2026)

Workshop Organizers

Cecília Coelho	Helmut Schmidt University, Germany
M. Fernanda P. Costa	University of Minho, Portugal

Luís Ferrás	University of Minho, Portugal
Eliana Costa e Silva	Polytechnic of Porto, Portugal

Workshop Program Committee Members

Osvaldo Gramaxo de Freitas	Universitat de València, Spain
Estela Bicho	University of Minho, Portugal
Flora Ferreira	University of Porto, Portugal
Bruno Veloso	Universidade do Porto, Portugal
Luís Louro	University of Minho, Portugal
Weronika Wojtak	CCG/ZGDV Institute, Portugal
Wolfram Erlhagen	University of Minho, Portugal

6th Workshop on Privacy in the Cloud/Edge/IoT World 2026 (PCEIoT 2026)

Workshop Organizers

Lelio Campanile	Università degli Studi della Campania Luigi Vanvitelli, Italy
Mauro Iacono	Università degli Studi della Campania Luigi Vanvitelli, Italy
Michele Mastroianni	University of Foggia, Italy

Workshop Program Committee Members

Fiammetta Marulli	University of Campania Luigi Vanvitelli, Italy
Christian Riccio	Università degli Studi della Campania Luigi Vanvitelli, Italy
Arcangelo Castiglione	University of Salerno, Italy
Gianni D'Angelo	University of Salerno, Italy
Biagio Boi	University of Salerno, Italy
Luigi Piero Di Bonito	Università degli Studi della Campania, Italy

Preserving Our Past: Spatial and Remote Sensing Technologies for Cultural Heritage in a Changing Climate (POP 2026)

Workshop Organizers

Maria Danese	Institute of Heritage Science of the National Research Council of Italy (CNR ISPC), Italy
Rosa Lasaponara	Institute of Methodologies for Environmental Analysis (CNR IMAA), Italy
Nicola Masini	Institute of Heritage Science of the National Research Council of Italy (CNR ISPC), Italy

Workshop Program Committee Members

Dario Gioia	Institute of Heritage Science of the National Research Council of Italy (CNR ISPC), Italy
Marilisa Biscione	Institute of Heritage Science of the National Research Council of Italy (CNR ISPC), Italy
Valentina Florio	CNR, Italy
Antonio Minervino Amodio	Institute of Heritage Science of the National Research Council of Italy (CNR ISPC), Italy

Scientific Computing Infrastructure (SCI 2026)

Workshop Organizers

Vladimir Korkhov	Saint Petersburg State University, Russia
Elena Stankova	Saint Petersburg State University, Russia

Workshop Program Committee Members

Adam Belloum	University of Amsterdam, The Netherlands
Dmitrii Vasiunin	Deutsche Telekom Cloud Services E.P.E., Greece
Serob Balyan	Institute for Informatics and Automation Problems, National Academy of Sciences of Armenia, Armenia
Suren Abrahamyan	Institute for Informatics and Automation Problems of NAS RA, Armenia
Ashot Gevorkyan	Institute for Informatics and Automation Problems NAS of Armenia, Armenia

Michal Hnatič	P.J. Šafárik University in Košice, Slovak Republic
Martin Vala	Pavol Jozef Šafárik University, Slovak Republic
Nodir Zaynalov	Samarkand Branch of Tashkent University of Information Technologies Named After Muhammad al Khwarizmi, Uzbekistan
Michail Panteleyev	Saint Petersburg Electrotechnical University, Russia
Alexander Degtyarev	Saint Petersburg State University, Russia
Alexander Bogdanov	St. Petersburg State University, Russia
Nataliia Kulabukhova	AutoTech, Russia
Alexander Degtyarev	St.Petersburg State University, Russia
Alexander Bogdanov	St. Petersburg State University, Russia

Socio-Economic and Environmental Models for Land Use Management (SEMLUM 2026)

Workshop Organizers

Debora Anelli	Sapienza University of Rome, Italy
Pierluigi Morano	Polytechnic University of Bari, Italy
Francesco Tajani	Sapienza University of Rome, Italy
Felicia Di Liddo	Polytechnic University of Bari, Italy

Workshop Program Committee Members

Benedetto Manganelli	Università degli Studi della Basilicata, Italy
Emma Sabatelli	Sapienza Università di Roma, Italy
María Paz Sáez-Pérez	University of Granada, Spain
Yasmine Selim	Ain Shams University, Egypt
Daniela Tavano	University of Calabria, Italy
Francesco Calabrò	Mediterranea University of Reggio Calabria, Italy
Pierfrancesco De Paola	University of Naples Federico II, Italy
Francesco Paolo Del Giudice	University of Rome La Sapienza, Italy
Paola Amoruso	LUM Giuseppe Degennaro University, Italy
Rubina Canesi	Università degli Studi di Padova, Italy
Daniele La Rosa	University of Catania, Italy
Giuseppe Cerullo	Sapienza University of Rome, Italy
Antonella Roma	Sapienza University of Rome, Italy
Francesca Fariello	Sapienza University of Rome, Italy

Ports and logistics of the future - Smartness and Sustainability (SmartPorts 2026)

Workshop Organizers

Andrea Gallo	Università degli Studi di Trieste, Italy
Giuseppe Borruso	University of Trieste, Italy
Gianfranco Fancello	University of Cagliari, Italy
Veronica Camerada	Università degli studi di Sassari, Italy
Clara Di Fazio	University of Trieste, Italy

Workshop Program Committee Members

Patrizia Serra	University of Cagliari, Italy
Francesco Bruzzone	Università Iuav di Venezia, Italy
Marco Mazzarino	University Iuav of Venice, Italy
Maria del Mar Munoz	University of Cádiz, Spain
Ainhoa Amaro	EVOCONS, Spain
Brunella Brundu	University of Sassari, Italy
Giovanni Mauro	University of Campania Luigi Vanvitelli, Italy
Maria Ronza	University of Naples Federico II, Italy
Marcello Tadini	Università del Piemonte Orientale, Italy
Luca Toneatti	University of Trieste, Italy
Mara Ladu	University of Cagliari, Italy
Ginevra Balletto	University of Cagliari, Italy
Francesca Sinatra	University of Trieste, Italy

Sustainable evolution of long-Distance frEight and paSsenger Transport (SOLIDEST 2026)

Workshop Organizers

Antonio Russo	University of Enna Kore, Italy
Giuseppe Musolino	Università Mediterranea di Reggio Calabria, Italy
Socrates Basbas	Aristotle University of Thessaloniki, Greece
Oana Dinu	National University of Science and Technology Politehnica Bucharest, Romania
Sanja Šurdonja	University of Rijeka, Croatia
Elen Twrdy	University of Ljubljana, Slovenia

Workshop Program Committee Members

Tiziana Campisi	University of Enna Kore, Italy
Massimo Di Gangi	Università di Messina, Italy
Antonio Polimeni	Università di Messina, Italy
Efstathios Bouhouras	Aristotle University of Thessaloniki, Greece
Florin Rusca	National University of Science and Technology Politehnica Bucharest, Romania
Vittorio Astarita	University of Calabria, Italy
Giuseppe Guido	Università della Calabria, Italy
Stefanos Tsigdinos	National Technical University of Athens, Greece
Chiara Garau	University of Cagliari, Italy

Sustainability Performance Assessment: Models, Approaches, and Applications Toward Interdisciplinary and Integrated Solutions (SPA 2026)

Workshop Organizers

Francesco Scorza	University of Basilicata, Italy
Sabrina Lai	University of Cagliari, Italy
Jolanta Dvarioniene	Kaunas University of Technology, Lithuania
Michele Campagna	University of Cagliari, Italy

Workshop Program Committee Members

Corrado Zoppi	University of Cagliari, Italy
Maria Cerreta	University of Naples Federico II, Italy
Giuseppe Borruso	University of Trieste, Italy
Chiara Garau	University of Cagliari, Italy
Ginevra Balletto	University of Cagliari, Italy
Alenka Poplin	Iowa State University, USA
Francesco Zullo	University of L'Aquila, Italy
Alessandro Marucci	University of L'Aquila, Italy
Federica Leone	University of Cagliari, Italy
Simone Corrado	University of Basilicata, Italy
Alfonso Annunziata	University of Basilicata, Italy
Rachele Vanessa Gatto	Università della Basilicata, Italy
Francesco Paolo Del Giudice	University of Basilicata, Italy

Specifics of smart cities development in Europe (SPEED 2026)

Workshop Organizers

Chiara Garau	University of Cagliari, Italy
Katarína Vitálišová	Matej Bel University, Slovak Republic
Marco Fanfani	University of Florence, Italy
Anna Vaňová	Matej Bel University, Slovak Republic
Kamila Borsekova	Matej Bel University, Slovak Republic
Paola Zamperlin	University of Pisa, Italy

Workshop Program Committee Members

Francesca Maltinti	University of Cagliari, Italy
Silvia Rossetti	Università degli Studi di Parma, Italy
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Floriana Zucaro	University of Naples Federico II, Italy
Giovanna Concu	University of Cagliari, Italy
Paolo Nesi	University of Florence, Italy
Pierfrancesco Bellini	University of Florence, Italy
Gerardo Carpentieri	University of Naples Federico II, Italy
Radoslav Kožiak	Matej Bel University, Slovak Republic
Lucia Dobrotová	Matej Bel University in Banská Bystrica, Slovak Republic
Veronika Piscová	Institute of Landscape Ecology Slovak Academy of Sciences, Slovak Republic
Darina Rojíková	Matej Bel University, Slovak Republic
Katarína Sýkorová	Matej Bel University, Slovak Republic
Silvia Ondrejičková	Comenius University Bratislava, Slovak Republic
Vladimír Ondrejička	Comenius University Bratislava, Slovak Republic
Maros Finka	Comenius University Bratislava, Slovak Republic
Jan Nevima	Masaryk University, Czech Republic
Christian Corral Burau	Leibniz University Hannover, Germany
Kamila Turečková	Silesian University in Opava, Czech Republic
Luciano Alessandro Ipsaro Palesi	University of Florence, Italy
Enrico Collini	University of Florence, Italy
Gianni Pantaleo	University of Florence, Italy
Krishnamoorthy Ranganathan	Chennai Institute of Technology, India

Smart, Safe, and Health Cities (SSHC 2026)

Workshop Organizers

Chiara Garau	University of Cagliari, Italy
Gerardo Carpentieri	University of Naples Federico II, Italy
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Martina Carra	University of Brescia, Italy
Vincenza Torrisi	University of Catania, Italy
Silvia Rossetti	Università degli Studi di Parma, Italy

Workshop Program Committee Members

Romano Fistola	University of Naples Federico II, Italy
Francesca Maltinti	University of Cagliari, Italy
Barbara Caselli	Università degli Studi di Parma, Italy
Tiziana Campisi	University of Enna Kore, Italy
Katarína Sýkorová	Matej Bel University, Slovak Republic
Tazyeen Alam	Università degli Studi di Cagliari, Italy
Francesco Zullo	University of L'Aquila, Italy
Giovanna Concu	University of Cagliari, Italy
Anastasia Stratigea	National Technical University of Athens, Greece
Ezio Riggi	National Research Council, Italy
Miriam Distefano	CNR, Italy

Smart and Sustainable Island Communities (SSIC 2026)

Workshop Organizers

Chiara Garau	University of Cagliari, Italy
Anastasia Stratigea	National Technical University of Athens, Greece
Yiota Theodora	National Technical University of Athens, Greece
Giovanna Concu	University of Cagliari, Italy
Tiziana Campisi	University of Enna Kore, Italy
Tanja Congiu	University of Sassari, Italy

Workshop Program Committee Members

Francesca Maltinti	University of Cagliari, Italy
Mauro Coni	University of Cagliari, Italy
Paola Zamperlin	University of Florence, Italy

Tarek Teba	University of Portsmouth, UK
Gerardo Carpentieri	University of Naples Federico II, Italy
Floriana Zucaro	University of Naples Federico II, Italy
Roberto Minunno	Curtin University, Australia
Apostolos Lagarias	University of Thessaly, Greece
Giuseppina Vacca	University of Cagliari, Italy
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Tazyeen Alam	Università degli Studi di Cagliari, Italy
Mauro D'Apuzzo	University of Cassino and Southern Lazio, Italy
Mana Dastoum	Polytechnic University of Madrid, Spain
Alfonso Annunziata	University of Basilicata, Italy
Barbara Caselli	Università degli Studi di Parma, Italy
Silvia Rossetti	Università degli Studi di Parma, Italy
Gian Felice Giaccu	University of Sassari, Italy
Anastasia Stratigea	National Technical University of Athens, Greece
Vincenza Torrisi	University of Catania, Italy

Sustainable Tourism Economy, Silk Road, Reuse and Enhancement of Cultural Heritage (STESilkRoad 2026)

Workshop Organizers

Celestina Fazia	Università di Enna Kore, Italy
Alexandru Ilies	University of Oradea, Romania
Dorina Camelia Ilies	University of Oradea, Romania
Cristina Natoli	Ministero dell'Ambiente e della Sicurezza Energetica, Italy
Bahodirhon Safarov	Samarkand Branch of Tashkent State University of Economics, Uzbekistan
Clara Stella Vicari Aversa	Università degli Studi Mediterranea di Reggio Calabria, Italy

Workshop Program Committee Members

Giulia Fernanda Grazia Catania	Università di Enna Kore, Italy
Federica Sortino	Università degli Studi di Enna Kore, Italy
Thowayeb Hassan	King Faisal University, Saudi Arabia
Tatar Corina	University of Oradea, Romania
Maria Francesca Faro	Ministry of Education and Merit (MIM), Italy

Theoretical and Computational Chemistry and its Applications (TCCM 2026)

Workshop Organizers

Noelia Faginas Lago	Università di Perugia, Italy
Andrea Lombardi	Università di Perugia, Italy
Marcos Mandado Alonso	University of Vigo, Spain

Workshop Program Committee Members

Jose Manuel Hermida Ramon	Universidade de Vigo, Spain
Luca Mancini	University of Perugia, Italy
Stefano Falcinelli	University of Perugia, Italy
Ángel Martín Pendás	Universidad de Oviedo, Spain
Enrique M. Cabaleiro Lago	Universidade de Santiago de Compostela, Spain
Giacomo Giorgi	Università degli Studi di Perugia, Italy
Stefano Evangelisti	Université de Toulouse, France
Nicolás Ramos Berdullas	University of Vigo, Spain
Juan José Nogueira	Universidad Autónoma de Madrid, Spain
Giacomo Pannacci	Università degli Studi di Perugia, Italy

Transport Infrastructures for Smart Cities (TISC 2026)

Workshop Organizers

Francesca Maltinti	University of Cagliari, Italy
Mauro Coni	University of Cagliari, Italy
Chiara Garau	University of Cagliari, Italy
Benedetto Barabino	University of Brescia, Italy
Nicoletta Rassu	Università di Cagliari, Italy
James Rombi	University of Cagliari, Italy

Workshop Program Committee Members

Tiziana Campisi	University of Enna Kore, Italy
Mauro D'Apuzzo	University of Cassino and Southern Lazio, Italy
Roberto Ventura	Università degli Studi di Brescia, Italy
Roberto Minunno	Curtin University, Australia

Giovanna Concu	University of Cagliari, Italy
Claudia Loggia	University of KwaZulu-Natal, South Africa

Transforming Urban Analytics: The Impact of Crowdsourced Mapping and Advanced AI Techniques on Future Cities (Tr-UrbAna 2026)

Workshop Organizers

Ayse Giz Gulnerman Gengec	Ankara Hacı Bayram Veli University, Turkey
Müslüm Hacar	Yildiz Technical University, Turkey

Workshop Program Committee Members

Beniamino Murgante	University of Basilicata, Italy
Abdulkadir Memduhoğlu	Harran University, Turkey
Zeynel Abidin Polat	Izmir Katip Celebi University, Turkey
Güzide Miray Perihanoglu	Van Yuzuncu Yil University, Turkey
Tugba Memisoglu Baykal	Hacı Bayram Veli University, Turkey
Federico Mara	University of Pisa, Italy
Himmet Karaman	Istanbul Technical University, Turkey
Özge Öztürk Hacar	Yildiz Technical University, Turkey

From structural to TRAnsformative-change of City Environment: challenges & solutions & perspectives (TRACE 2026)

Workshop Organizers

Maria Rosaria Guarini	Sapienza University of Rome, Italy
Pierluigi Morano	Polytechnic University of Bari, Italy
Francesco Sica	Sapienza University of Rome, Italy
Alejandro Segura	Universidad Politécnica de Madrid, Spain
Yilsy María Nuñez Guerrero	Universidad Politécnica de Madrid, Spain

Workshop Program Committee Members

Maria Cerreta	University of Naples Federico II, Italy
Francesco Scorza	University of Basilicata, Italy

Francesco Calabrò	Mediterranea University of Reggio Calabria, Italy
Felicia Di Liddo	Polytechnic University of Bari, Italy
Marco Locurcio	Polytechnic University of Bari, Italy
Valeria Saiu	University of Cagliari, Italy
Lucia Della Spina	University Mediterranea of Reggio Calabria, Italy
Fabrizio Battisti	University of Florence, Italy
Giuseppe Cerullo	Sapienza University of Rome, Italy
Fabiana Forte	Università della Campania Luigi Vanvitelli, Italy
Marta Dell'Ovo	Politecnico di Milano, Italy
Cristina Coscia	Politecnico di Torino, Italy
Debora Anelli	Sapienza University of Rome, Italy

Urban Artificial Intelligence for Modelling, Analysis & Planning (Urban plAnnIng 2026)

Workshop Organizers

Chiara Garau	University of Cagliari, Italy
Romano Fistola	University of Naples Federico II, Italy
Beniamino Murgante	University of Basilicata, Italy
Dessislava Petrova Antonova	Sofia University, GATE Institute, Bulgaria
Gianni Pantaleo	University of Florence, Italy
Rosa Anna La Rocca	University of Naples Federico II

Workshop Program Committee Members

Paola Zamperlin	University of Florence, Italy
Francesco Zullo	University of L'Aquila, Italy
Gerardo Carpentieri	University of Naples Federico II, Italy
Luciano Alessandro Ipsaro Palesi	University of Florence, Italy
Tazyeen Alam	Università degli Studi di Cagliari, Italy
Reza Askarizad	University of Cagliari, Italy
Mana Dastoum	Polytechnic University of Madrid, Spain
Alfonso Annunziata	University of Basilicata, Italy

Urban Space Accessibility and Mobilities (USAM 2026)

Workshop Organizers

Chiara Garau	University of Cagliari, Italy
Alessandro Plaisant	University of Sassari, Italy
Barbara Caselli	Università degli Studi di Parma, Italy
Mauro D'Apuzzo	University of Cassino and Southern Lazio, Italy
Francesco Pinna	University of Cagliari, Italy
Silvio Nocera	IUAV University of Venice, Italy

Workshop Program Committee Members

Mauro Coni	University of Cagliari, Italy
Martina Carra	University of Brescia, Italy
Tiziana Campisi	University of Enna Kore, Italy
Reza Askarizad	University of Cagliari, Italy
Giuseppe Cappelli	University of Cassino and Southern Lazio, Italy
Floriana Zucaro	University of Naples Federico II, Italy
Gerardo Carpentieri	University of Naples Federico II, Italy
Youssef El Ganadi	International University of Rabat, Morocco
Yiota Theodora	National Technical Univeristy of Athens, Greece
Silvia Rossetti	Università degli Studi di Parma, Italy
Aime Tsinda	University of Rwanda, Rwanda
Paola Zamperlin	University of Florence, Italy
Luciano Alessandro Ipsaro Palesi	University of Florence, Italy
Tazyeen Alam	Università degli Studi di Cagliari, Italy
Mariano Pernetti	Università della Campania Luigi Vanvitelli, Italy
Salvatore Leonardi	University of Catania, Italy
Francesco Galante	University of Naples Federico II, Italy

Virtual Reality, Augmented Reality and Applications (VRA 2026)

Workshop Organizers

Damiano Perri	University of Perugia, Italy
Osvaldo Gervasi	University of Perugia, Italy
Paolo Nesi	University of Florence, Italy

Workshop Program Committee Members

Taihoon Kim	Chonnam National University, South Korea
JungYoon Kim	Gachon University, South Korea
David Berti	ART SpA, Italy

Workshop on Advanced and Computational Methods for Earth Science applications (WACM4ES 2026)

Workshop Organizers

Luca Piroddi	University of Cagliari, Italy
Patrizia Caprizzi	University of Palermo, Italy
Marilena Cozzolino	Università del Molise, Italy
Giuseppina Vacca	University of Cagliari, Italy
Luigi Capozzoli	Hydrogeosite Laboratory of CNR-IMAA, Italy
Nikos Papadopoulos	Foundation for Research and Technology Hellas, Greece

Workshop Program Committee Members

Sebastiano D'Amico	University of Malta, Malta
Enrica Vecchi	University of Cagliari, Italy
Donatella Rita Fiorino	Università degli Studi di Cagliari, Italy
Chiara Garau	University of Cagliari, Italy
Emanuele Colica	University of Malta, Malta
Raffaele Martorana	Università di Palermo, Italy
Raffaele Persico	Università della Calabria, Italy
Andrea Angelini	Institute of Heritage Science of the National Research Council of Italy (CNR ISPC), Italy
Paolo Mauriello	Università del Molise, Italy
Dedalo Marchetti	Italian National Institute of Geophysics and Volcanology, Italy
Veronica Pazzi	University of Florence, Italy
Agnese Innocenti	University of Florence, Italy
Maurizio Ercoli	University of Perugia, Italy
Luciano Galone	University of Malta, Malta
Amr Abd El-Raouf	Zagazig University, Egypt
Peter Iregbeyen	University of Malta, Malta
Pietro Milillo	University of Houston, USA
Tommaso Beni	University of Florence, Italy

Cristiano Fidani	Central Italy Electromagnetic Network, Italy
Deodato Tapete	Italian Space Agency (ASI), Italy
Enzo Rizzo	University of Ferrara, Italy
Guido Luzi	Centre Tecnològic de Telecomunicacions de Catalunya, Spain
Sina Saneiyan	Binghamton University, USA
Alex Sendrós	Universitat de Barcelona, Spain
Mercedes Solla	Universidade de Vigo, Spain
Alessandra Carollo	University of Palermo, Italy
Andrea Vergnano	Università di Torino, Italy
Giorgio Alaia	Università di Perugia, Italy
Giorgio De Donno	Sapienza University of Rome, Italy
Samuel Bignardi	University of Chieti and Pescara, Italy
Athos Agapiou	Cyprus University of Technology, Cyprus
Gregory Tsokas	Aristotle University of Thessaloniki, Greece
Alexandru Hegyi	Oslo University, Norway
Ilaria Barone	Università degli Studi di Padova, Italy
Yuan Qi	University of Hong Kong, China
Carmen Cuenca García	Universitat de València, Spain
Ilaria Catapano	Institute for Electromagnetic Sensing of the Environment (IREA) CNR, Italy
Mariano Lisi	University of Basilicata, Italy
Adrien Dimech	École de Technologie Supérieure, Canada
Davide Campo	Independent Researcher, UK
Chiara Torre	University of Malta, Malta
Anna Forte	University of Bologna, Italy
Adam Gauci	University of Malta, Malta
Vittoria Borghese	TNO, The Netherlands
Andrea Scianna	Institute for High Performance Computing and Networking (ICAR) CNR, Italy
Maria Antonia Brovelli	Politecnico di Milano, Italy
Pietro Milillo	University of Houston, USA
Gregory Tsokas	Aristotle University of Thessaloniki, Greece

Sponsoring Organizations

ICCSA 2026 would not have been possible without tremendous support from many organizations and institutions, for which all organizers and participants of ICCSA 2026 express their sincere gratitude:



Universidade do Minho
Escola de Engenharia

University of Minho, Portugal (https://www.uminho.pt/)



University of Perugia, Italy (https://www.unipg.it)



University of Basilicata, Italy (http://www.unibas.it)

University of Cagliari, Italy (https://unica.it/)

Monash University, Australia (https://www.monash.edu/)

Kyushu Sangyo University, Japan (https://www.kyusan-u.ac.jp/)

Venue

ICCSA 2026 took place at: **the University of Minho campus:**
Campus de Gualtar
4710-057 Braga
Portugal





Additional Reviewers

Amr Abd Elraouf	Zagazig University, Egypt
Lidia Aceto	Università del Piemonte Orientale, Italy
Ana Maria Acu	Lucian Blaga University of Sibiu, Romania
Vera Afreixo	University of Aveiro, Portugal
Akif Akgul	Hitit University, Turkey
Sania Akhtar	Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Pakistan
Giorgio Alaia	Università di Perugia, Italy
Tazyeen Alam	Università degli Studi di Cagliari, Italy
Arnulfo Alanis	Tecnológico Nacional de México, Mexico
Sandra Aleixo	University of Lisbon, Portugal
Earl Ryan Aleluya	Mindanao State University Iligan Institute of Technology, Philippines
Muhammad Junaid Ali	Université de Haute Alsace, France
Filipe Alvelos	Universidade do Minho, Portugal
Claudia Amadei	University of Padua, Italy
Paulo Amorim	Escola de Matemática Aplicada da Fundação Getulio Vargas, Brazil
Ana Paula Amorim	Universidade do Minho, Portugal
Paola Amoruso	LUM Giuseppe Degennaro University, Italy
Debora Anelli	Sapienza University of Rome, Italy
Mariarosaria Angrisano	Pegaso University, Italy
Alfonso Annunziata	University of Basilicata, Italy
Chiara Anselmi	Università di Pisa, Italy
Shigeyoshi Aoyama	KEK, Japan
Tatsumi Aoyama	University of Tokyo, Japan
Bernady O. Apduhan	Kyushu Sangyo University, Japan
Pablo Luiz Araujo Munhoz	Federal University of Ouro Preto, Brazil
Muhammad Waqas Arshad	University of Bologna, Italy
Daniela Ascenzi	Università degli Studi di Trento, Italy
Laura Ascione	Università degli Studi di Napoli Federico II, Italy
Reza Askarizad	University of Cagliari, Italy
Vanessa Assumma	University of Bologna, Italy
Vittorio Astarita	University of Calabria, Italy
Maria Fernanda Augusto	BITrum Research Center, Spain
Asja Aulisio	Politecnico di Torino, Italy
Samat Baigereyev	D. Serikbayev East Kazakhstan Technical University, Kazakhstan
Marco Baioletti	University of Perugia, Italy
Sara Balderas Díaz	University of Cádiz, Spain

Ginevra Balletto	University of Cagliari, Italy
Serob Balyan	Institute for Informatics and Automation Problems, National Academy of Sciences of Armenia, Armenia
Benedetto Barabino	University of Brescia, Italy
Sebastiano Barbieri	Politecnico di Torino, Italy
Ilaria Barone	Università degli Studi di Padova, Italy
Alice Barreca	Politecnico di Torino, Italy
Socrates Basbas	Aristotle University of Thessaloniki, Greece
Rosaria Battarra	Consiglio Nazionale delle Ricerche, Italy
Silvia Battino	University of Sassari, Italy
Fabrizio Battisti	University of Florence, Italy
Patricia Beites	University of Beira Interior, Portugal
David Berti	Art Spa, Italy
Samuel Bignardi	University of Chieti and Pescara, Italy
Francesco Bimbo	University of Foggia, Italy
Francesca Biondi	Politecnico di Torino, Italy
Marilisa Biscione	CNR, Italy
Alexander Bogdanov	St. Petersburg State University, Russia
Biagio Boi	University of Salerno, Italy
Alessandro Bombini	Istituto Nazionale di Fisica Nucleare, Italy
Ana Borges	Instituto Politécnico do Porto, Portugal
Vittoria Borghese	TNO, The Netherlands
Giuseppe Borruso	University of Trieste, Italy
Kamila Borsekova	Matej Bel University, Slovak Republic
Efstathios Bouhouras	Aristotle University of Thessaloniki, Greece
Ana Cristina Braga	University of Minho, Portugal
Diego Brandão	Federal Center for Technological Education of Rio de Janeiro, Brazil
Maria Antonia Brovelli	Politecnico di Milano, Italy
Francesco Bruzzone	Università Iuav di Venezia, Italy
Francesca Buglione	CNR ISPC, Italy
Enrique M. Cabaleiro Lago	Universidade de Santiago de Compostela, Spain
Francesco Calabrò	Mediterranea University of Reggio Calabria, Italy
Rogerio Calazan	Instituto de Estudos do Mar Almirante Paulo Moreira - Brazilian Navy, Brazil
Lelio Campanile	Università degli Studi della Campania Luigi Vanvitelli, Italy
Tiziana Campisi	University of Enna Kore, Italy
Dario Campisi	University of Perugia, Italy
Rubina Canesi	Università degli Studi di Padova, Italy
Alessandra Capolupo	Polytechnic University of Bari, Italy

Caterina Caprioli	Politecnico di Torino, Italy
Dario Caramiello	University of Naples Parthenope, Italy
Gerardo Carpentieri	University of Naples Federico II, Italy
Martina Carra	University of Brescia, Italy
Barbara Caselli	Università degli Studi di Parma, Italy
Danny Casprini	Politecnico di Milano, Italy
Arcangelo Castiglione	University of Salerno, Italy
Giulia Fernanda Grazia Catania	Università di Enna Kore, Italy
Ilaria Catapano	CNR IREA, Italy
Laura Cavalcante	Instituto Superior de Engenharia do Porto, Portugal
Giulio Cavana	Politecnico di Torino, Italy
Valeria Cera	University of Naples Federico II, Italy
Maria Cerreta	University of Naples Federico II, Italy
Giuseppe Cerullo	Sapienza University of Rome, Italy
Daniele Cesini	CNAF, INFN, Italy
Kai Cheng	Kyushu Sangyo University, Japan
Giuseppe Ciciriello	University Of Naples Federico II, Italy
Birol Ciloglugil	Ege University, Turkey
Cecília Coelho	Helmut Schmidt University, Germany
Enrico Collini	University of Florence, Italy
Luís Conceicao	Instituto Superior de Engenharia do Porto, Portugal
Giovanna Concu	University of Cagliari, Italy
Tatar Corina	University of Oradea, Romania
Alejandro Cornejo	Universitat Politècnica de Catalunya, Spain
Simone Corrado	University of Basilicata, Italy
Florbela Correia	Polytechnic Institute of Viana do Castelo, Portugal
Franco Corti	University of Padua, Italy
Cristina Coscia	Politecnico di Torino, Italy
Lino Costa	University of Minho, Portugal
Eliana da Costa e Silva	Polytechnic Institute of Porto, Portugal
Alessandro Costantini	INFN, Italy
Alessia Crisopulli	Politecnico di Milano, Italy
Valerio Cutini	University of Pisa, Italy
Mauro D'Apuzzo	University of Cassino and Southern Lazio, Italy
Hiroshi Daisaka	Hitotsubashi University, Japan
Sebastiano Damico	University of Malta, Malta
Maria Danese	CNR ISPC, Italy
Gianni Dangelo	University of Salerno, Italy
Mana Dastoum	Polytechnic University of Madrid, Spain

Giulia Datola	Politecnico di Milano, Italy
Regina De Almeida	University of Trás-os-Montes and Alto Douro, Portugal
Elena De Angelis	Politecnico di Torino, Italy
Elise de Doncker	Western Michigan University, USA
Pasquale De Luca	Parthenope University of Naples, Italy
Luiza De Macedo Mourelle	State University of Rio de Janeiro, Brazil
Pierfrancesco De Paola	University of Naples Federico II, Italy
Alexander Degtyarev	St.Petersburg State University, Russia
Francesco Paolo Del Giudice	University of Basilicata, Italy
Piera Della Morte	Pegaso Telematic University, Italy
Federico Dellanna	Politecnico di Torino, Italy
Marta Dellovo	Politecnico di Milano, Italy
Federica Deri	Università di Pisa, Italy
Francesco Di Capua	University of Basilicata, Italy
Clara Di Fazio	University of Trieste, Italy
Gabriella Di Genova	Università degli Studi di Perugia, Italy
Benedetta Di Leo	Università Politecnica delle Marche, Italy
Felicia Di Liddo	Polytechnic University of Bari, Italy
Joana Dias	University of Coimbra, Portugal
Oana Dinu	National University of Science and Technology Politehnica Bucharest, Romania
Lucia Dobrotová	Matej Bel University in Banská Bystrica, Slovak Republic
Edda Donati	University of Rome Tor Vergata, Italy
Aziz Dursun	Virginia Tech University, USA
Zahra Ebadi Ansaroudi	Fondazione Bruno Kessler, Italy
Maurizio Ercoli	University of Perugia, Italy
Dario Esposito	Polytechnic University of Bari, Italy
Noelia Faginas Lago	Università di Perugia, Italy
Maria Irene Falcao	University of Minho, Portugal
Stefano Falcinelli	University of Perugia, Italy
Marco Fanfani	University of Florence, Italy
Marcos Farias	Instituto de Engenharia Nuclear, Brazil
Francesca Fariello	Sapienza University of Rome, Italy
Maria Francesca Faro	Ministry of Education and Merit (MIM), Italy
Eslam Farsimadan	University of Salerno, Italy
Sara Favargiotti	University of Trento, Italy
Celestina Fazia	Università di Enna Kore, Italy
Maksims Feofilovs	Riga Technical University, Latvia
Florbela P. Fernandes Cedri	Instituto Politecnico de Bragança, Portugal
Paula Odete Fernandes	Instituto Politecnico de Bragança, Portugal

Flora Ferreira	University of Porto, Portugal
Maddalena Ferretti	Università Politecnica delle Marche, Italy
Luís Ferrás	University of Porto, Portugal
Massimo Ficco	University of Salerno, Italy
Maros Finka	Masaryk University, Slovak Republic
Ugo Fiore	University of Salerno, Italy
Lorena Fiorini	University of L'Aquila, Italy
Romano Fistola	University of Naples Federico II, Italy
Beatriz Flamia Azevedo	Instituto Politecnico de Bragança, Portugal
Valentina Florio	CNR, Italy
Ali Forooghi	University of Campania Luigi Vanvitelli, Italy
Anna Forte	University of Bologna, Italy
Grigorios Fountas	Aristotle University of Thessaloniki, Greece
Andreas Fricke	Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany
Luigi Fusco Girard	Pegaso Telematic University, Italy
Francesco Galante	University of Naples Federico II, Italy
Stefano Galassi	University of Florence, Italy
Andrea Gallo	Università degli Studi di Trieste, Italy
Chiara Garau	University of Cagliari, Italy
Gustavo Gatica	Universidad Andrés Bello, Chile
Rachele Vanessa Gatto	Università della Basilicata, Italy
Ivan Gerace	Università degli Studi di Perugia, Italy
Amal Gerges	University of Cagliari, Italy
Osvaldo Gervasi	University of Perugia, Italy
Tasnim Gharaibeh	Kalamazoo College, USA
Gian Felice Giaccu	University of Sassari, Italy
Dario Gioia	CNR ISPC, Italy
Luca Giommi	INFN CNAF, Italy
Valentina Giorgetti	Università degli Studi di Perugial, Italy
Giacomo Giorgi	Università degli Studi di Perugia, Italy
Hiago Mayk Gomes De Araújo Rocha	National Laboratory for Scientific Computing, Brazil
Omar Gomez	Universidad Nacional de Colombia, Colombia
A. Manuela Goncalves	University of Minho, Portugal
Joaquim Goncalves	Polytechnic Institute of Cávado and Ave, Portugal
Pedro Henrique Gonzalez	Federal University of Rio de Janeiro, Brazil
Yusuke Gotoh	Okayama University, Japan
Osvaldo Gramaxo De Freitas	Universitat de València, Spain
Antonia Gravagnuolo	National Research Council Italy (CNR) Institute of Heritage Science, Italy

Teresa Guarda	Universidad Estatal Península De Santa Elena, Ecuador
Maria Rosaria Guarini	Sapienza University of Rome, Italy
Raphael Guerra	Universidade Federal Fluminense, Brazil
Hélia Guerra	University of the Azores, Portugal
Gabriel Guerrero Contreras	University of Cádiz, Spain
Carmen Guida	Università degli Studi di Napoli Federico II, Italy
Eligius Hendrix	Universidad de Málaga, Spain
Jose Manuel Hermida Ramon	Universidade de Vigo, Spain
Antonino Iannuzzo	Universitá degli Studi del Sannio, Italy
Luciano Alessandro Ipsaro Palesi	University of Florence, Italy
Issaku Kanamori	Riken, Japan
Yeliz Karaca	UMass Chan Medical School and Massachusetts Institute of Technology, USA
Himmet Karaman	Istanbul Technical University, Turkey
Tatyana Khachkova	IPGG SB RAS, Russia
Jungyoon Kim	Gachon University, South Korea
Vladimir Korkhov	Saint Petersburg State University, Russia
Iuliia Kozlova	University of Bologna, Italy
Radoslav KožIak	Matej Bel University, Slovak Republic
Ludovica La Rocca	University of Naples Federico II, Italy
Rosa Anna La Rocca	University of Naples Federico II, Italy
Adnane Labbaci	Ibn Tofail University, Morocco
Sabrina Lai	University of Cagliari, Italy
Giuseppe Francesco Cesare Lama	University of Naples Federico Ii, Italy
Isabella Maria Lami	Politecnico di Torino, Italy
Salvatore Leonardi	University of Catania, Italy
Federica Leone	University of Cagliari, Italy
Vita Leonessa	University of Basilicata, Italy
Barbara Lino	Università degli Studi di Palermo, Italy
Vadim Lisitsa	Institute of Petroleum Geology and Geophysics SB RAS, Russia
Marco Locurcio	Polytechnic University of Bari, Italy
Andrea Lombardi	Università di Perugia, Italy
Cristina Lopes	Polytechnic Institute of Porto, Portugal
Isabel Lopes	Instituto Politécnico de Bragança, Portugal, Italy
Rui Pedro Lopes	Instituto Politécnico de Bragança, Portugal
Nicolas Lori	University of the Azores, Portugal
Luís Louro	University of Minho, Portugal
Jader Lugon Junior	Instituto Federal Fluminense, Brazil
Guido Luzi	Centre Tecnològic de Telecomunicacions de Catalunya, Spain

Antonio S. Macias	University of A Coruña, Spain
Daniela Maiullari	Delft University of Technology, The Netherlands
Robert Makin	Western Michigan University, USA
Giorgia Malavasi	Politecnico di Torino, Italy
Francesca Maltinti	University of Cagliari, Italy
Luca Mancini	University of Perugia, Italy
Marcos Mandado Alonso	University of Vigo, Spain
Benedetto Manganelli	Università degli Studi della Basilicata, Italy
Federico Mara	University of Pisa, Italy
Livia Marcellino	University of Naples Parthenope, Italy
Dedalo Marchetti	INGV, Italy
Francesco Marchetti	University of Padua, Italy
Alessandra Marra	University of Salerno, Italy
Barbara Martelli	INFN, Italy
Valerio Martinelli	Università degli Studi di Napoli Federico II, Italy
Ángel Martín Pendás	Universidad de Oviedo, Spain
Alessandra Mascitelli	University of Chieti Pescara, Italy
Michele Mastroianni	University of Foggia, Italy
Luís Miguel Matos	Universidade do Minho, Portugal
Hideo Matsufuru	High Energy Accelerator Research Organization, Japan
Luis Mazon	BITrum Research Group, Spain
Beatrice Mecca	Politecnico di Torino, Italy
Valeria Mele	University of Naples Federico II, Italy
Abdulkadir Memduhoğlu	Harran University, Turkey
Joao Mendes	Instituto Politecnico de Braganca, Portugal
Luis Mendes Gomes	University of the Azores, Portugal
Paolo Mengoni	Hong Kong Baptist University, China
Daniela Menna	Università degli Studi della Campania Luigi Vanvitelli, Italy
Antonio Minervino Amodio	CNR ISPC, Italy
Roberto Minunno	Curtin University, Australia
Fernando Miranda	Universidade do Minho, Portugal
Andrea Montanino	International Centre for Numerical Methods in Engineering, Spain
Andrea Montanino	University of Naples Parthenope, Italy
Ana Monteiro	University of Coimbra, Portugal
Leila Moradi	University of Salerno, Italy
Joao Morais	ITAM, Mexico
Filipe Mota Pinto	Instituto Politécnico de Leiria, Portugal
Ricardo Moura	NOVA University of Lisbon, Portugal

Maria Mourao	Polytechnic Institute of Viana do Castelo, Portugal
Eugenio Muccio	University of Naples Federico II, Italy
Beniamino Murgante	University of Basilicata, Italy
Giuseppe Musolino	Università Mediterranea di Reggio Calabria, Italy
Naohito Nakasato	University of Aizu, Japan
Isabel Natário	Nova University of Lisbon, Portugal
Nadia Nedjah	State University of Rio de Janeiro, Brazil
Grazia Neglia	Pegaso Telematic University, Italy
Paolo Nesi	University of Florence, Italy
Keigo Nitadori	Riken, Japan
Juan José Nogueira	Universidad Autónoma de Madrid, Spain
Mikhail Novikov	Sobolev Institute of Mathematics SB RAS, Russia
Yilsy María Nuñez Guerrero	Universidad Politécnica de Madrid, Spain
Ana Oliveira	Academia Militar, Portugal
Carlo Olivieri	Pegaso University, Italy
Vladimír Ondrejička	Comenius University in Bratislava, Slovak Republic
Beniamino Onorati	Università degli Studi della Basilicata, Italy
Fabian Bernal Orozco Barrera	UNAM, Mexico
M. Fernanda P. Costa	University of Minho, Portugal
António Pacheco	Instituto Superior Técnico, Portugal
Emerson Paiva	University of Minho, Portugal
Simone Palladino	Università degli Studi della Campania L. Vanvitelli, Italy
Francesco Palmieri	University of Salerno, Italy
Simona Panaro	University of Sussex Business School, UK
Claudia Panciera	IRSA CNR, Italy
Giacomo Pannacci	Università degli Studi di Perugia, Italy
Gianni Pantaleo	University of Florence, Italy
Eric Pardede	La Trobe University, Australia
Rosaria Parente	Universitas Mercatorum, Italy
Vittorio Paris	University of Bergamo, Italy
Luca Parodi	University of Genoa, Italy
Rui Pascoal	University of Coimbra, Portugal
Hugo Peixoto	University of Minho, Portugal
Francesco Peluso	University of Naples Parthenope, Italy
Ana Pereira	Polytechnic Institute of Bragança, Portugal
Amina Pereno	Politecnico di Torino, Italy
Martin Perez Perez	University of Vigo, Spain
Donato Perfetto	University of Campania Luigi Vanvitelli, Italy
Güzide Miray Perihanoglu	Van Yuzuncu Yil University, Turkey

Mariano Pernetti	Università della Campania Luigi Vanvitelli, Italy
Damiano Perri	University of Perugia, Italy
Pasquale Perrini	Consorzio Interuniversitario per l'Idrologia, Italy
Raffaele Persico	Università della Calabria, Italy
Lucia Petronio	University of Bologna, Italy
Rossella Petti	Politecnico di Bari, Italy
Francesco Picariello	Universitas Mercatorum, Italy
Angela Pilogallo	IMAA CNR, Italy
Natalia Pingaro	Politecnico di Milano, Italy
Telmo Miguel Pires Pinto	University of Minho, Portugal
Veronika Piscová	Institute of Landscape Ecology, Slovak Academy of Sciences, Slovak Republic
Zeynel Abidin Polat	Izmir Katip Celebi University, Turkey
Giuliano Poli	Università di Napoli Federico II, Italy
Antonio Polimeni	Università di Messina, Italy
Maurizio Pollino	ENEA, Italy
Mohammad Pourfouladi	Politecnico di Milano, Italy
Dmitry Prokhorov	IPGG SB RAS, Russia
Faiza Qamar	National University of Sciences and Technology, Pakistan
Yuan Qi	University of Hong Kong, China
Siyu Qiu	University of Bologna, Italy
Arunasalam Rahunanthan	Central State University, USA
Joelmir Ramos	Universidade do Estado do Rio de Janeiro, Brazil
Bruna Ramos	University of Minho, Portugal
Nicolás Ramos Berdullas	University of Vigo, Spain
Krishnamoorthy Ranganathan	Chennai Institute of Technology, India
Nicoletta Rassu	Università di Cagliari, Italy
Roberta Ravanelli	University of Liège, Belgium
Michela Ravanelli	Sapienza Università di Roma, Italy
Manuela Rebaudengo	Politecnico di Torino, Italy
Simone Rebegoldi	University of Modena and Reggio Emilia, Italy
Pier Francesco Recchi	Università di Napoli Federico II, Italy
Rommel Regis	Saint Joseph's University, USA
Marco Reis	University of Coimbra, Portugal
James Rhodes	Western Michigan University, USA
Luigi Maciel Ribeiro	State University of Rio de Janeiro, Brazil
Corrado Rindone	Università degli Studi Mediterranea di Reggio Calabria, Italy
Ana Maria A. C. Rocha	Universidade do Minho, Portugal
Humberto Rocha	University of Coimbra, Portugal
Carlos Rodrigues	Instituto Politécnico de Bragança, Portugal

Ivan Rodriguez Conde	University of Arkansas at Little Rock, USA
Darina Rojíková	Matej Bel University, Slovak Republic
Diana Rolando	Politecnico di Torino, Italy
Antonella Roma	Sapienza University of Rome, Italy
James Rombi	University of Cagliari, Italy
Elisabetta Ronchieri	INFN CNAF, Italy
Marzio Rosi	University of Perugia, Italy
Silvia Rossetti	Università degli Studi di Parma, Italy
Marco Rossitti	Politecnico di Milano, Italy
Florin Rusca	National University of Science and Technology Politehnica Bucharest, Romania
Yvonne Russo	University of Campania L. Vanvitelli, Italy
Antonio Russo	University of Enna Kore, Italy
Ana Sofia Rézio	Polytechnic Institute of Lisbon, Portugal
Emma Sabatelli	Sapienza Università di Roma, Italy
Matteo Sacchet	University of Turin, Italy
Sabrina Sacco	University of Naples Federico II, Italy
Lucia Saganeiti	Institute of Methodologies for Environmental Analysis, National Research Council, Italy
Valeria Saiu	University of Cagliari, Italy
Francesco Santini	University of Perugia, Italy
Rafael Santos	National Institute for Space Research (INPE), Brazil
Edcarllos Santos	Universidade Federal do Sul da Bahia, Brazil
Valentino Santucci	Università per Stranieri di Perugia, Italy
Andrea Sarcina	Università di Padova, Italy
Ilaria Scarcelli	Politecnico di Bari, Italy
Andrea Scianna	National Research Council of Italy, Italy
Francesco Scorza	University of Basilicata, Italy
Ester Scotto Di Perta	University of Naples Federico II, Italy
Alejandro Segura	Universidad Politécnica de Madrid, Spain
Yasmine Selim	Ain Shams University, Egypt
Ines Sena Cedri	Instituto Politécnico de Bragança and University of Minho, Portugal
Alex Sendrós	Universitat de Barcelona, Spain
Ricardo Severino	University of Minho, Portugal
Mojtaba Sharifzadeh	Università di Trento, Italy
Luigi Sibille	Norwegian University of Science and Technology, Norway
Francesco Sica	Sapienza University of Rome, Italy
Adriano Silva	Instituto Politecnico de Bragança, Portugal
Angela Silva	Universidade do Minho, Portugal

Paula Simões	Military Academy, Portuguese Army, Portugal
Francesca Sinatra	University of Trieste, Italy
Mercedes Solla	Universidade de Vigo, Spain
Sergey Solovyev	Sobolev Institute of Mathematics SB RAS, Russia
Maria Somma	University of Naples Federico II, Italy
Federica Sortino	Università degli Studi di Enna Kore, Italy
Eleonora Sottile	University of Cagliari, Italy
Martina Spinelli	Politecnico di Torino, Italy
Elena Stankova	Saint Petersburg State University, Russia
Alexandra Stankulova	Politecnico di Torino, Italy
Giorgia Sugoni	Links Foundation, Italy
Katarína Sýkorová	Matej Bel University, Slovak Republic
Francesco Tajani	Sapienza University of Rome, Italy
Deodato Tapete	Italian Space Agency (ASI), Italy
Rodrigo Tapia-McClung	Centro de Investigación en Ciencias de Información Geoespacial, Mexico
Maria Tartari	CNR, Italy
Daniela Tavano	University of Calabria, Italy
Tarek Teba	University of Portsmouth, UK
Ana Paula Teixeira	Universidade do Minho, Portugal
Wagner Telles	Universidade Federal Fluminense, Brazil
Luis Teodoro	University of Oslo, Norway
Maria Filomena Teodoro	Instituto Superior Técnico, Portugal
Yiota Theodora	National Technical University of Athens, Greece
Elena Todella	University of Parma, Italy
Graca Tomaz	Instituto Politécnico da Guarda, Portugal
Chiara Torre	University of Malta, Malta
Francesca Torrieri	Politecnico di Milano, Italy
Vincenza Torrisi	University of Catania, Italy
Arianna Travaglini	Lucian Blaga University of Sibiu, Romania
Stefanos Tsigdinos	National Technical University of Athens, Greece
Kamila Turečková	Silesian University in Opava, Czech Republic
Elen Twrdy	University of Ljubljana, Slovenia
Toshihiro Uchibayashi	Kyushu University, Japan
Takahiro Ueda	Juntendo University, Japan
Generoso Vaiano	Pegaso University, Italy
Aleksandar Valjarević	University of Belgrade, Serbia
Gianmarco Vanuzzo	University of Perugia, Italy
Dmitrii Vasiunin	Deutsche Telekom Cloud Services E.P.E., Greece
Mohammad Vazir Panah	University of Salerno, Italy
Anna Vaňová	Matej Bel University, Slovak Republic
Enrica Vecchi	University of Cagliari, Italy

Roberto Ventura	Università degli Studi di Brescia, Italy
Ferdinando Verardi	Pegaso Telematic University, Italy
Clara Stella Vicari Aversa	Università degli Studi Mediterranea di Reggio Calabria, Italy
Gianluca Vinti	Università degli Studi di Perugia, Italy
Katarína Vitálišová	Matej Bel University, Slovak Republic
Weronika Wojtak	CCG/ZGDV Institute, Portugal
Fukuko Yuasa	KEK, Japan
Madina Yussubaliyeva	D. Serikbayev East Kazakhstan Technical University, Kazakhstan
Paola Zamperlin	University of Florence, Italy
Francisco Zayas Gato	Universidade da Coruña, Spain
Nodir Zaynalov	Samarkand Branch of Tashkent University of Information Technologies Named After Muhammad Al Khwarizmi, Uzbekistan
Ljiljana Zivkovic	Republic Geodetic Authority, Serbia
Piero Zizzania	University of Naples Federico II, Italy
Renato Zona	Università della Campania L. Vanvitelli, Italy
Floriana Zucaro	University of Naples Federico II, Italy
Francesco Zullo	University of L'Aquila, Italy
Giovanni Zurlo	INFN CNAF, Italy
Özge Öztürk Hacar	Yildiz Technical University, Turkey
Isidora Đurić	University of Novi Sad, Serbia

Plenary Lectures

Digital Twins for Cities – The Rise of the Socio-Technical System

Claudia van der Laag

Department of Built Environment, Oslo Metropolitan University, Norway

Abstract. Cities are entering a new era in which digital twins are no longer simply technical replicas of urban systems, but evolving socio-technical infrastructures that shape how cities are planned, governed, and experienced. Across the world, urban digital twins are being adopted to address some of the most pressing challenges of our time — climate adaptation, urban resilience, mobility transitions, decarbonisation, infrastructure management, and inclusive governance. Yet as these technologies become increasingly sophisticated, a fundamental question emerges: can digital twins truly support more democratic and human-centred cities, or will they reinforce existing structures of power and expertise? In this keynote, Claudia van der Laag explores the transformation of urban digital twins from data-driven modelling tools into collaborative platforms for urban decision-making and societal change. It argues that the future of digital twins lies not only in advances in artificial intelligence, real-time data integration, simulation, and predictive analytics, but equally in the ability to connect technological innovation with governance, participation, and human agency. Drawing on emerging international research and practical case studies, the keynote examines the complex ecosystem surrounding urban digital twins, bringing together planners, policymakers, researchers, technologists, and communities. It highlights how digital twins increasingly function as meeting points between multiple forms of knowledge — technical, political, social, and experiential — and why interdisciplinary collaboration is essential for their success. The talk critically reflects on the tensions between technological ambition and civic inclusion, exploring themes such as transparency, interoperability, participation, and trust. It also discusses how digital twins can become powerful environments for scenario exploration, urban experimentation, and collective learning, enabling cities to better navigate uncertainty and complexity in an age of rapid urban transformation. Ultimately, this keynote presents a vision of digital twins not as neutral technological systems, but as living socio-technical infrastructures capable of reshaping urban governance itself. The challenge ahead is therefore not only how accurately cities can be modelled, but how inclusively they can be imagined, negotiated, and governed through these emerging digital environments.

Prescriptive AI in Practice: Bridging Optimisation, Generative AI and Agentic Decision-Making

Bernardo Almada-Lobo

Universidade do Porto, Faculdade de Engenharia, INESC TEC, Portugal

Abstract. Prescriptive AI, viewed through the lens of mathematical optimisation and operations research, addresses complex business decision-making problems involving limited resources, multiple decision options and competing objectives. This talk considers the full decision pipeline: problem framing, model formulation, computational solution, implementation, monitoring and organisational adoption. The practical success of prescriptive AI depends less on algorithmic sophistication alone than on the quality of the translation between real-world decision contexts and formal optimisation models. In practice, this translation remains a major bottleneck. Objectives are often ambiguous, constraints are incomplete or tacit, data may be fragmented, and the operational consequences of a recommendation may be poorly understood. As a result, optimisation projects can fail even when the underlying mathematical model is technically sound. The talk will address the persistent gap between optimisation theory and deployment practice. While optimisation has generated substantial value in domains such as supply chain planning, logistics, production, revenue management, workforce planning and inventory control, many initiatives still face a significant risk of under-delivery. Common failure modes include incorrect problem framing, weak data readiness, brittle assumptions, infeasible or non-actionable recommendations, limited explainability, insufficient ownership by business users and inadequate post-deployment monitoring. Closing this gap requires moving from one-off model building to an operational lifecycle in which models are stress-tested, embedded in workflows, monitored for drift and revised as the business environment changes. Recent developments in generative and agentic AI create new opportunities to reduce these barriers. Generative AI can support the elicitation of decision-maker preferences, the structuring of unstructured operational knowledge, the drafting of candidate formulations and the explanation of model outputs in business language. Agentic AI can extend this role by orchestrating multi-step workflows that involve model formulation, solver-code generation, execution, debugging, infeasibility diagnosis, sensitivity analysis and human review. However, these capabilities should not be interpreted as replacing optimisation expertise. Rather, they create a new interface between human judgement, analytical modelling and computational decision support. This talk will discuss a cross-fertilisation agenda between optimisation, generative AI and agentic AI, with a focus on how these technologies should be calibrated to the decision at hand.

Net-Zero Communities through Life Cycle Thinking: Canadian Lessons and Global Implications

Kasun Hewage

School of Engineering (Okanagan Campus), Clean Energy Research Centre (CERC), University of British Columbia (UBC), Canada

Abstract. This keynote presentation discusses how life cycle thinking supports the transformation of communities toward sustainable, resilient, and net-zero futures. Drawing on Canadian research, practice, and policy experience, the presentation highlights how integrated life cycle approaches can move beyond conventional carbon accounting to inform practical decision-making across buildings, infrastructure, energy systems, waste management, water systems, and industrial processes [1–8]. The keynote features research conducted at the Life Cycle Management Laboratory (LCML) at the University of British Columbia, Canada, an internationally recognized research hub led by Kasun Hewage. LCML brings together engineers, scientists, policy researchers, postdoctoral fellows, visiting scholars, and graduate researchers to translate data-intensive sustainability research into actionable climate solutions. The laboratory's work is organized around six interconnected research themes: energy systems, urban development, water systems, waste management, construction and asset management, and industrial processes and products. Together, these themes provide an integrated platform for advancing life cycle–based pathways to net-zero communities. Supported by strong partnerships with government, industry, and community stakeholders, LCML has contributed to a substantial body of applied sustainability research, including more than 400 journal articles, 100 conference papers, and numerous technical reports that have informed real-world planning and implementation. This keynote presentation highlights selected success stories that demonstrate how life cycle thinking can be translated into practical strategies for reducing environmental impacts while strengthening community resilience. These insights are especially relevant to the interdisciplinary focus of ICCSA 2026.

References

1. Weerasinghe, O., Hewage, K., Razi, F., Fatmi, M., Sadiq, R.: Applicability of land use models for sustainable urban planning: a review. J. Urban Plann. Dev. ASCE, **151**(3) (2025). https://doi:10.1061/JUPDDM.UPENG-5705

2. Sultana, S. R., et al.: Looking beyond energy efficiency targets: life cycle sustainability of mechanical and water heating equipment in Canadian homes. J. Clean. Prod. Elsevier, **473**, 143469 (2024). https://doi.org/10.1016/j.jclepro.2024.143469
3. Liyanage, D. R., Hewage, K., Hussain, S. H., Razi, F., Sadiq, R.: Climate adaptation of existing buildings: a critical review on planning energy retrofit strategies for future climate. Renew. Sustain. Energy Rev. Elsevier, **199**, 114476 (2024). https://doi.org/10.1016/j.rser.2024.114476
4. Liyanage, D. R., Hewage, K., Karunathilake, H., Sadiq, R.: Feasibility of carbon-capturing in building heating systems: a life cycle thinking-based approach. International J. Greenhouse Gas Control. Elsevier, **132** (2023). https://doi.org/10.1016/j.ijggc.2023.104056
5. Moudgil, V., Sadiq, R., Bakhtavar, E., Paudel, A., Hewage, K.: A cloud-oriented data-analysis framework to analyze peak demand dynamics in institutional building clusters. Sustain. Cities Soci. Elsevier, **111**, 105553 (2024). https://doi.org/10.1016/j.scs.2024.105553
6. Liyanage, D. R., Hewage, K., Ghobadi, M., Sadiq, R.: Thermal resiliency of single-family housing stock under extreme hot and cold conditions. Energy Build. Elsevier, **323**, 114809 (2024). https://doi.org/10.1016/j.enbuild.2024.114809
7. Wijayasekera, S. C., et al.: Sustainability of waste-to-hydrogen conversion pathways: a life cycle thinking-based assessment. Energy Convers. Manage. Elsevier, **270** (2022). https://doi.org/10.1016/j.enconman.2022.116218
8. Jayasena, A., Hewage, K., Siddiqui, O., Sadiq, R.: Socio-economic and environmental cost-benefit analysis of passive houses: a life cycle perspective. J. Clean. Prod. Elsevier, **373** (2022). https://doi.org/10.1016/j.jclepro.2022.133718

Machine Learning Based Methodologies for Grape and Wine Production

Pedro Melo-Pinto

University of Trás-os-Montes and Alto Douro, Engineering Department, Centre for the Research and Technology of Agro-Environmental and Biological Sciences, Portugal

Abstract. Wine production is a major area in the Portuguese economy. The sector accounts for approximately 2.6% of Portugal's GDP, and is crucial to the development of inland regions. Wine quality depends on grape quality, which is a result of different factors such as harvesting at the optimal point of maturity, or selecting the grapes according to quality features. The growing importance of precision agriculture demands novel methods to make the most out of the data generated. Machine learning approaches are particularly suitable for modelling complex processes, where too many arguments are involved. Many of the agricultural processes such as yield estimation or ripening fall into this category. However, several challenges may arise: even if models are available for particular cases, general models are difficult to come by, largely due to the high variability involved; with such models, there is a need for huge amounts of data, in a way which will not always be easy to obtain. This is particularly true in the case of grape production and the wine industry, where the number of grape varieties, climate, soil, etc., introduce a degree of variability that is uncommon in other sectors. One way of obtaining more information in such cases has been the use of images, whether RGB, multi-spectral, or hyperspectral. The use of images introduces additional challenges, such as data size and lighting conditions, which underscores the need for data-driven modelling methodologies. The use of machine learning (ML) techniques combined with images has been establishing itself as a viable alternative for predicting yields or quality parameters in grapes, and assisting with critical harvesting decisions. Several ML approaches have been proposed to handle such data characteristics, but selecting a suitable methodology that best addresses the problem under study and making sure it generalizes well is, most of the time, a cumbersome task. We will address three wine and grape production problems, yield estimation, grape quality assessment, and illness detection, presenting different difficulties. Yield estimation is a major aspect in the wine production industry. The number of berries is one of the most relevant yield aspects, and can be estimated using computer vision, namely image segmentation algorithms. Lighting is one of the key factors in computer vision approaches, and occlusions—which are common in this problem—can be minimized, but require estimation models to achieve reliable results. Early illness detection is, of course, of the utmost importance, with an impact on production and quality. Detecting early symptoms and distinguishing them from other conditions is a difficult task, which makes the use of machine learning algorithms advisable, using automatic visual inspection of the leaves, addressed as a pixel-based classification problem. However, the available datasets are limited and

complex to compile, as different diseases typically require very specific time-frames and locations. The use of generative models to overcome this obstacle, in conjunction with robust classification models, leads to models that are intended to be as general as possible. The evaluation of the grape's maturation profile based on the evolution of their oenological parameters over time is one of the most important aspects for ensuring wine quality, playing a decisive role in the definition of the optimal time for harvesting. The use of hyperspectral imaging combined with learning algorithms represents a faster and non-destructive grape ripeness assessment method. We focus in two fundamental aspects to address the resulting data variability: a possible wavelength bands selection (with the purpose of reducing the dimensionality of data without losing predictive power) and the generalization ability of the ML model under such demanding conditions.

Contents

Computational Methods, Statistics and Industrial Mathematics (CMSIM 2026)

Determinants of Bachelor's Degree Completion: The Role of Pre-University Background and Academic Progress 3
Francesca Pierri and Chrys Caroni

Dual-Branch Hidden Markov Model for Rolling Bearing Degradation Analysis 20
Indrawata Wardhana, Amal Gouiaa-Mtibaa, Pascal Vrignat, and Frédéric Kratz

Research on the Gimballed Inertial Navigation System Based on Mathematical Description of Platform Motion 37
Olha Sushchenko

Adaptive Sampling Design for Bayesian Spatial Risk Mapping of Maritime Surveillance in Central Portugal 51
Belchior Miguel, Isabel Natário, Rui Gonçalves de Deus, and Paula Simões

Investigating of the Operation of a Railgun System with Laboratory Proportion 69
P. Sampaio Pereira, Mário Gatta, and M. Filomena Teodoro

Determinants of Housing Prices in Portugal: A Municipal-Level Analysis 87
Maria J. Ferreira, Aldina Correia, Vanda Lima, and Diogo Ribeiro

Modeling Risk in HIV Care: A Data-Driven Approach to Patient Profiling 99
M. Filomena Teodoro and Teresa Oliveira

The Influence of Fatigue on Decision-Making in Military Context 115
Francisco Abreu, Nuno Almeida, Renato Santos, Rui Lucena, and Paula Simões

A Framework for the Statistical Analysis of Fixed Effects Models Using *R*, Illustrated with an Example in Biometry 132
Teresa A. Oliveira and Maria Filomena Teodoro

Computational Optimization and Applications (COA 2026)

Entropy Based Estimation of Risk Neutral Density Dynamics 149
Ana Fernandes, Rui Pascoal, and Ana M. Monteiro

Predicting Consumer Acceptance of Gamification Strategies: A Machine Learning Approach 167
Patrícia Santos, Cristela Bairrada, and Joana Dias

Multi-objective Optimization of Distributed Job Shop Scheduling Problems with Customer Time Requirements 182
Francisco dos Santos, Lino Costa, and Leonilde Varela

Optimizing Directional Change Trading Strategies with Adaptive Thresholds for the Brazilian Stock Market 199
Lucas E. Mieri and Sanderson L. Gonzaga de Oliveira

Optimization and Technical-Economic Evaluation of DC Microgrids with Integrated Renewable Energy Sources 215
Adelaide Cerveira, Paulo Silva, and José Baptista

(In)approximability of the Maximum 2-Clique Problem for Chordal Graphs . . . 231
Yuichi Asahiro, Eiji Miyano, and Shunta Nakamura

Evaluating AI-Generated Surrogate Optimization Algorithms for Expensive Black-Box Optimization 247
Rommel G. Regis

Heuristic Approaches for Personnel Scheduling in a Call Center with Social Fairness Considerations 266
Anaís Veloso Silva, Telmo Pinto, and Cláudio Alves

Distributionally Robust Optimisation for Prepositioning Wildfire Fighting Resources 283
Agostinho Agra, Filipe Alvelos, Francisco Marques, Ana Raquel Xambre, and Helena Alvelos

A Convergence Analysis of a 1-Norm Neighborhood Mehrotra-Type Algorithm 301
A. Teixeira and R. De Almeida

Mixed-Integer Linear Programming for Production Scheduling with Robot Allocation in a Textile Company 314
Maria Coelho Lima, Carina Pimentel, and Filipe Alvelos

On the Use of Range Volatility Estimators for Rough Fractional Stochastic Volatility Model 332
Joana Sendas, Rui Pascoal, and Humberto Rocha

A Reinforcement Learning–Guided General Variable Neighborhood Search for the Traveling Salesman Problem 351
Karina Assini Andreatta, Maria Claudia Silva Boeres, and Renato Elias Nunes de Moraes

Computational Methods in Advanced Networking Technologies (CoMANT 2026)

When Heterogeneous Modalities Synchronize: Mitigating Desynchronization Attacks in Multimodal Online Federated Learning 371
Kainat Ibrar, Asim Ibrahim, and Francesco Palmieri

A Fractional Epidemic Model for Multi-malware Attacks in Wireless Sensor Networks 387
Eslam Farsimadan, Leila Moradi, Marcello Trovati, and Francesco Palmieri

Sensitivity Analysis of the Multi Malware Attacks Model in Wireless Sensor Networks 406
Leila Moradi, Eslam Farsimadan, Marcello Trovati, and Francesco Palmieri

Computational Astrochemistry 2026 (CompAstro 2026)

From N(^{2}D) Back to N_2: A Theoretical Investigation of the N(^{2}D) + $CH_2NH \rightarrow N_2 + CH_3$ Reaction and Implications for the Atmospheric Chemistry of Titan 425
Luca Mancini, Marco Parriani, Marzio Rosi, and Nadia Balucani

Production and Characterization of Cyanoacetylene Dications by Synchrotron Radiation 436
Marco Parriani, Franco Vecchiocattivi, Fernando Pirani, Michele Alagia, Robert Richter, Stefano Stranges, Raimund Feifel, Simonetta Cavalli, and Stefano Falcinelli

Reactions of N(^{2}D) with Polycyclic Aromatic Hydrocarbons: A Computational Approach 448
Marzio Rosi, Nadia Balucani, Luca Mancini, Giacomo Pannacci, Marco Parriani, Dimitrios Skouteris, and Gianmarco Vanuzzo

Can OSiS Be Formed in the Interstellar Medium? A Computational Investigation of the OH + SiS Reaction with a Kinetic Analysis 461
Nadia Balucani, Luca Mancini, Dimitrios Skouteris, Cecilia Ceccarelli, Linda Podio, Claudio Codella, Gabriella Di Genova, and Marzio Rosi

Workshop on Computational Science and HPC (CSHPC 2026)

Numerical Evaluation of 3-Loop "Baseball" and "Running Shirt" Feynman Diagrams Using GPU-Accelerated QMC Integration and Extrapolation, Based on Sector Decomposition 477
Elise de Doncker, Hiroshi Daisaka, Tadashi Ishikawa, Kiyoshi Kato, Naohito Nakasato, and Fukuko Yuasa

Development and Validation of the Computational Code GEM for Simulating Quantum Few-Body Systems through Optimization on the Supercomputer Fugaku 492
Shigeyoshi Aoyama, Yusuke Tsunoda, Noritaka Shimizu, Hideo Matsufuru, Issaku Kanamori, and Emiko Hiyama

Acceleration of Multicomponent Multiple-Precision Arithmetic with Branch-Free Algorithms and SIMD Vectorization 508
Tomonori Kouya

Advection of Smooth and Non-smooth Tracers Using Spectral Bicubic Interpolation 525
Takeshi Enomoto

Matrix Functions Computation in Hardware Accelerators 542
Daniel Oliveira, Juan Acebrón, and José Monteiro

Cities, Technologies and Planning (CTP 2026)

Data (Re)conciliation: Audit and Harmonization Pipelines for Municipal Geospatial Data 563
Rodrigo João Bastos Mendes, Ertuğrul Doğruluk, Alexandre Valente Conceição Pereira de Sousa, and Pedro Correia Cravo Pimenta

Integrating Renewable Energy Communities into Historic Urban Landscapes: A Framework for Sustainable Conservation 577
Mara Ladu, Ginevra Balletto, Alessandra Milesi, Federica Isola, and Emilio Ghiani

How Do Local and Regional Strategies Address Blue-Green Infrastructure? the Case of Slovakia 592
Ondrej Mitaľ, Michaela Lukačínová, and Silvia Ručinská

The Challenges of Urban Regeneration Between Conservation and Transformation in Coastal Areas. The Case Study of La Caletta, Sardinia, Italy 610
Ginevra Balletto and Mara Ladu

Machine Learning Methods for Fire Risk Assessment in Wildland Urban Interfaces - Case Study Attica Region, Greece 622
Angeliki Papazoglou, Apostolos Lagarias, and Anastasia Stratigea

SmartParcel System: A Geospatial Platform for Enhancing Cross-Fertilisation Between Spatial Planning and Territorial Cohesion Policies in Serbia 641
Ljiljana Živković

Author Index 659

Computational Methods, Statistics and Industrial Mathematics (CMSIM 2026)

Determinants of Bachelor's Degree Completion: The Role of Pre-University Background and Academic Progress

Francesca Pierri[1(✉)] and Chrys Caroni[2]

[1] Department of Economics, University of Perugia, Perugia, Italy
francesca.pierri@unipg.it
[2] Department of Mathematics, National Technical University of Athens, Athens, Greece
ccar@math.ntua.gr

Abstract. Universities can improve their training systems by systematically analyzing students' academic trajectories to better understand progression patterns and identify early risk factors for dropout. Administrative data—such as exams taken, credits earned, grades, and time to completion—provide valuable information for this purpose.

In this study, a cohort of students enrolled in three-year degree programs at the University of Perugia was followed over a ten-year period (2015–2025), exceeding the nominal duration of the degree in order to capture final academic outcomes. Multinomial logistic regression and competing risks models based on the Fine–Gray approach were applied to identify the main determinants of academic success and dropout.

The results show that, while pre-university background characteristics are relevant at entry, their predictive power decreases over time. Academic performance indicators—particularly accumulated and remaining credits—emerge as the main drivers of both graduation and attrition. The comparison between models highlights a trade-off between predictive accuracy and early interpretability, with time-to-event approaches providing a more appropriate representation of student trajectories.

These findings support the importance of early monitoring of academic performance to identify at-risk students and design targeted support strategies, contributing to more effective and data-driven university policies.

Keywords: university dropout · logistic regression · survival analysis · competing risks

1 Introduction

The analysis of students' academic trajectories has become a central topic in higher education research, as universities increasingly rely on empirical evidence to improve teaching effectiveness, support student success, and reduce

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 3–19, 2027.
https://doi.org/10.1007/978-3-032-30521-3_1

dropout rates. Monitoring students' progress through administrative data—such as exams taken, credits accumulated, grades obtained, and time to completion—provides institutions with valuable insights into how students navigate their academic pathways. These data allow universities to identify patterns in performance, detect potential difficulties early in the academic career, and design targeted interventions aimed at improving persistence and degree completion.

Student persistence and graduation outcomes have long been studied within the broader literature on higher education. Foundational theoretical frameworks emphasize the role of both individual characteristics and institutional environments in shaping students' educational trajectories. In particular, Tinto's model of student integration highlights how academic and social integration within the university environment influence persistence decisions and the likelihood of completing a degree [19,20]. Similarly, Bean [1] proposed a model inspired by organizational turnover theory, emphasizing how students' attitudes, academic performance, and institutional experiences affect their decisions to remain enrolled. These theoretical contributions have provided the basis for extensive empirical research investigating the determinants of student success and dropout.

More recent studies confirm that university dropout remains a significant challenge for higher education systems worldwide. Recent reviews estimate that nearly one in five students does not complete a university degree, with substantial consequences for individuals, institutions, and society. Non-completion is associated with poorer labour market outcomes and increased economic vulnerability, highlighting the importance of identifying the factors that influence persistence and degree completion [14].

A substantial strand of literature focuses on the role of pre-university background in shaping academic outcomes. Socio-economic status, parental education, prior academic performance, and the type of secondary school attended are consistently identified as key determinants of student success [3,6,14]. Students from disadvantaged socio-economic backgrounds often face structural barriers that increase the probability of dropout or delay in completing their studies. Empirical evidence shows that lower family cultural and economic capital significantly increases the risk of dropout, suggesting that inequalities accumulated before university continue to influence outcomes during higher education [15].

In addition to background characteristics, research increasingly highlights the importance of academic progress during university as a key predictor of final outcomes. Indicators such as credit accumulation, failed examinations, grade point average, and course participation provide strong signals of students' academic trajectories. Beyond initial background characteristics, an increasing number of studies highlight the importance of academic progress during the early stages of university as a key predictor of long-term success. Early academic performance, credit accumulation, and timely completion of examinations are often strong indicators of eventual graduation [8,12]. Students who fall behind in the first years of their studies frequently face increasing difficulties in catching up, which can ultimately lead to prolonged enrollment or dropout. Moreover, recent predictive studies show that accumulated credits and failed courses are among the

most powerful predictors of university dropout and academic success, with their importance evolving throughout the academic career [21]. Early academic performance is therefore often considered a critical indicator for identifying students at risk of falling behind or abandoning their studies.

The growing availability of administrative and digital learning data has led to the emergence of new research fields such as learning analytics and educational data mining, which aim to transform educational data into actionable knowledge. These approaches apply statistical and machine learning techniques to analyze student behavior, predict academic outcomes, and support evidence-based decision-making in higher education institutions. Recent literature highlights the increasing use of predictive analytics and artificial intelligence to identify students at risk of dropout and to design early intervention strategies aimed at improving retention and academic success [9].

Within this context, longitudinal and event-history approaches have become particularly relevant for studying students' academic careers. Survival analysis and competing risks models allow researchers to analyze the timing of events such as graduation, dropout, or prolonged enrollment while accounting for the dynamic nature of academic pathways. These methods provide a more nuanced understanding of student progression and allow institutions to identify critical periods during which the risk of dropout is highest [7].

Building on this growing literature, the present study examines the determinants of bachelor's degree completion by analyzing the academic trajectories of students enrolled in three-year programs at the Faculty of Economics of the University of Perugia. Using administrative data tracking students from enrollment over a ten-year period (2015–2025), the analysis observes academic careers beyond the nominal duration of the degree in order to capture different possible outcomes, including timely graduation, delayed completion, or non-completion.

The empirical analysis applies two complementary methodological approaches: multinomial logistic regression and competing risks survival models. The first method allows us to classify students according to different academic outcomes and identify the factors associated with each trajectory. The second approach models the timing of events during the academic career while accounting for the presence of multiple possible exit states. By comparing these two methodologies, the study highlights their respective advantages in terms of interpretability and their usefulness for transforming administrative data into predictive knowledge.

Overall, the objective of the paper is twofold. First, we aim to identify the main determinants of bachelor's degree completion by considering both students' pre-university background and indicators of academic progress during university studies. Second, we compare two statistical approaches widely used in the analysis of educational data to evaluate their effectiveness in modeling academic trajectories and supporting data-driven policies in higher education. Through this analysis, the study contributes to the understanding of students' academic pathways and provides empirical evidence that may support universities in designing policies aimed at promoting academic success and timely graduation.

The remainder of this paper is organized as follows. Section 2 outlines the methodological framework and describes the statistical techniques employed in the analysis. Section 3 presents the empirical application, including a description of the dataset and the main results. Section 4 concludes with a discussion of the findings and their implications.

2 Methodology

Statistical and machine learning methods can help create predictive models that estimate the chances of graduating on time, graduating late, or dropping out. Some of the most popular methods include Logistic Regression, Decision Tree Learning, Random Forests, and Survival Analysis, which is used to predict the time until graduation or dropout.

Here multinomial logistic regression and competing risks models based on the Fine–Gray approach were applied to consider the different cause of dropping out. Descriptive statistics were used before developing the models. Model selection was performed on a training dataset comprising 80% of the observations, randomly drawn without replacement from the full sample. The fitted model was then evaluated on the remaining 20% of the data, serving as an out-of-sample validation set, to assess predictive performance and generalizability. This approach allowed us to quantify the model's accuracy and robustness, ensuring that the estimated relationships were not driven by overfitting to the training data.

All analyses were primarily conducted using SAS® software [17,18], which was employed for the estimation of multinomial logistic regression models and Fine–Gray competing risks models through the `PROC LOGISTIC` [17] and `PROC PHREG` [18] procedures. To complement the inferential analysis and to evaluate the predictive performance of the competing risks model, additional analyses were performed in the R statistical environment. In particular, the `timeROC` package [4,5] was employed to compute time-dependent receiver operating characteristic curves and corresponding area under the curve (AUC) measures.
This complementary use of R allowed for a more detailed assessment of model discrimination over time, which is not directly provided within the standard SAS implementation for competing risks models. In this way, SAS ensured the estimation of the regression framework, while R was used for advanced predictive performance evaluation and graphical diagnostics.

2.1 Multinomial Logistic Regression

Multinomial logistic regression is basically an extension of binary logistic regression that looks at how a set of explanatory variables relates to a categorical dependent variable with more than two unordered outcomes. It is part of the family of discrete choice models and works well when the response variable has several mutually exclusive options that do not follow any natural order. Thanks to its flexibility in dealing with nominal outcomes and multiple options, multinomial logistic regression is commonly used in areas like economics, sociology, political science, marketing, and education, especially for research on choice behavior,

career paths, or educational trajectories. Multinomial logistic regression with a nominal dependent variable can be modeled using the generalized logit (glogit) link, which represents the log-odds of each possible outcome compared to a chosen reference category.

Let Y be a categorical variable with J unordered categories ($j = 1, 2, ..., J$) and let $X = (X_1, X_2, ..., X_p)$ denote a vector of explanatory variables. By convention, one category (e.g., category J) is chosen as the baseline. The generalized logit model then specifies the log-odds of being in category j relative to the reference category J as a linear function of the predictors:

$$\ln \frac{P(Y = j \mid X)}{P(Y = J \mid X)} = \beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jp}X_p, \quad j = 1, \ldots, J-1 \tag{1}$$

From this, the predicted probabilities for each category are:

$$P(Y = j \mid X) = \frac{\exp(\beta_{j0} + \beta_{j1}X_1 + \cdots + \beta_{jp}X_p)}{1 + \sum_{k=1}^{J-1} \exp(\beta_{k0} + \beta_{k1}X_1 + \cdots + \beta_{kp}X_p)}, \quad j = 1, \ldots, J-1 \tag{2}$$

$$P(Y = J \mid X) = \frac{1}{1 + \sum_{k=1}^{J-1} \exp(\beta_{k0} + \beta_{k1}X_1 + \cdots + \beta_{kp}X_p)} \tag{3}$$

Here, each coefficient β_{jk} represents the effect of predictor X_k on the log-odds of being in category j relative to the reference category J. The model ensures that predicted probabilities are positive and sum to 1.

Parameters were estimated via maximum likelihood. Multicollinearity among covariates was assessed prior to model estimation. Variable selection was conducted using a stepwise procedure based on Wald statistics as an exploratory tool to identify a parsimonious set of predictors. Model selection was further supported by likelihood-based criteria, including the Akaike Information Criterion (AIC), the Schwarz Bayesian Criterion (SBC), and the -2 Log Likelihood, as provided by SAS output. Given the well-known limitations of stepwise procedures, including potential model instability and inflated Type I error rates, results were interpreted with caution and emphasis was placed on consistency across model specifications. Model fit was assessed using standard goodness-of-fit measures, including the Hosmer-Lemeshow test [13], while predictive performance was evaluated through the Brier Score.

2.2 Competing Risks Analysis

Survival models are often used in risk analysis to estimate how long it might take for one or more events to happen, for example death, failure of a machine, or relapse of a disease. This kind of study is also called time-to-event analysis or event history analysis. The survival function gives the probability that the event occurs after time t,

$$S(t) = P(T > t) \tag{4}$$

where T is the random variable representing the time to the event. The survival function is related to the cumulative distribution function $F(t)$, that represents the total chance that an event will happen within time t.

$$S(t) = 1 - F(t). \tag{5}$$

A particular interest lies in the hazard function, that is the instantaneous risk that the event occurs at time t, given survival until time t

$$h(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t | T > t]}{\delta t} = \frac{f(t)}{S(t)}\,. \tag{6}$$

Competing risks analysis deals with situations in which an observational unit is at risk of experiencing one of several mutually exclusive events, called competing events and the occurrence of one event precludes the occurrence of the others. Two main approaches exist in this context: the cause-specific hazard model and the subdistribution hazard model (Fine–Gray). The first analyzes the hazard of a particular event by treating other competing events as censored observations. We define the cause-specific hazard for the kth of K causes as

$$h_k(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t,\ D = k | T > t]}{\delta t}\,, \tag{7}$$

where D denotes the cause of failure. It models the instantaneous risk that the kth event occurs at time t, given that no event has occurred yet. However, it does not directly model the cumulative incidence of the event.

In this context we used the Fine–Gray approach [10,11,16] because it directly models the subdistribution hazard, which is linked to the cumulative incidence function (CIF) of a specific event in the presence of competing risks. In this case, individuals who have experienced one of the competing events remain in the risk set with special weighting rather than being censored. The subdistribution hazard function [10] estimates the risk of failure from cause k,

$$h_k^s(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t, D = k \mid T \geq t \text{ or } (T \leq t \text{ and } D \neq k)]}{\delta t} \tag{8}$$

and expresses the instantaneous rate of failure at time t among subjects that have not yet experienced an event of type k.

Model specification was guided by likelihood-based criteria (AIC) and by the magnitude and precision of subdistribution hazard ratios and their confidence intervals, with the aim of obtaining a parsimonious and interpretable model. The proportional hazards assumption was assessed for all covariates using standard diagnostic procedures. Evidence of non-proportionality was observed for the variable graduation mark. To address this issue, an interaction term between graduation mark and the logarithm of time was introduced and included in the model alongside the original graduation mark variable. This time-dependent specification allows the effect of graduation mark on the hazard to vary over time, providing a flexible adjustment for the non-proportionality while preserving interpretability of the covariate's effect on the outcome.

To assess model performance, both calibration and discrimination were evaluated. Calibration was examined through the comparison of cumulative incidence functions (CIFs) between training and validation samples. Discrimination was assessed by computing time-dependent area under the curve (AUC) measures within a competing risks framework. This approach allowed the evaluation of the model's ability to distinguish between different event types over time. Overall, this combined strategy provides a comprehensive assessment of model performance, integrating regression-based inference with predictive validation.

3 Application

3.1 Data

A cohort of students enrolled in 2015 in the three bachelor degree programs in Economics at the University of Perugia was followed for ten years. Each student was identified by an anonymous identification code. Along with the details collected during registration, such as gender, residence, secondary school attended, diploma, and graduation mark, information about their university career was also provided. Therefore, for each student, we have details like the exams they have taken, the training credits earned, the exam dates, the grades received, and their current status: graduated (status = 1), voluntary withdrawal from studies (status = 2), withdrawal due to inactivity (status = 3), or still active (status = 0).

Table 1. Distribution of students' current status by course

Degree Code		Status				
		0	1	2	3	Total
L021	N	24	209	79	64	376
L025	N	15	40	13	9	77
L026	N	3	19	6	13	41
Total	N	42	268	98	86	494
	%	8.5	54.25	19.84	17.41	100

The three bachelor degrees in Economics are located in different Umbria Region districts and are identified with codes: L021 (Business Administration in Perugia), L025 (Business Administration in Terni) and L026 (Business for Tourism in Assisi). The distribution of students by status and degree course (Table 1) shows that the highest percentage of students attends the L021 course (76.11%) (Table 3).
Moreover, keeping track of how long it takes students to finish their studies, we notice (Table 2) that just 54.25% completed their degree program. Among graduates, 37.31% finished within the expected duration of three years, 41.79%

Table 2. Distribution of students by status and time

		Time										
Status		1	2	3	4	5	6	7	8	9	10	Total
0	N	0	0	0	0	0	0	0	0	0	42	42
	%	0	0	0	0	0	0	0	0	0	8.5	8.5
1	N	0	0	100	112	38	8	6	3	1	0	268
	%	0	0	20.24	22.67	7.69	1.62	1.21	0.61	0.2	0	54.25
2	N	54	18	7	4	7	4	2	1	1	0	98
	%	10.93	3.64	1.42	0.81	1.42	0.81	0.4	0.2	0.2	0	19.84
3	N	0	0	2	2	7	3	63	5	4	0	86
	%	0	0	0.4	0.4	1.42	0.61	12.75	1.01	0.81	0	17.41
Total	N	54	18	109	118	52	15	71	9	6	42	494
	%	10.93	3.64	22.06	23.89	10.53	3.04	14.37	1.82	1.21	8.5	100

Table 3. Distribution of type of school attended by gender.

School		Female	Male	Total
		M	F	Total
Technical	N	120	86	206
	%	43.17	39.81	41.70
Scientific	N	103	43	146
	%	37.05	19.91	29.55
Classic	N	23	44	67
	%	8.27	20.37	13.56
Professional	N	20	18	38
	%	7.19	8.33	7.69
Mixed*	N	1	20	21
	%	0.36	9.26	4.25
Outside Italy	N	11	5	16
	%	3.96	2.31	3.24
Total	N	278	216	494

*Includes elements of classical and scientific

completed in four years, 14.18% in five years, while 6.72% required between six and ten years. Students belonging to status 2 (19.84%), closed their university position mainly in the first (55.10%) and second (18.37%) years. A further 17.41% of the students failed to complete their studies (status 3); they were removed for inactivity by the system mainly at the end of the seventh year (73.26%).

Students signing up for Economics courses live mainly in the Umbria region (86.23%) and attended a Technical (41.7%), Scientific (29.55%) or Classical (13.56%) high school.
The quantitative variables used in the analysis are shown along with their distributions in Table 4.

Table 4. Data distribution

Variable	N	Minimum	Maximum	Median	Mean	Std Dev
School Graduation Mark	492	0	100	76.00	76.30	16.37
Exam Grade Year 1 (mean 1)	404	18	30	24.00	23.90	3.10
Exam Grade Year 2	332	18	30	24.98	24.70	2.76
Exam Grade Year 3	318	18	30	25.89	25.48	2.59
Exam Grade Year 4	200	18	30	23.65	23.78	2.90
Exam Grade Year 5	72	18	30	23.00	23.23	3.30
Total exams completed	494	0	26	23.00	15.47	10.86
Total exams left	494	0	22	0.00	4.37	5.90
Exams completed year 1	494	0	10	4.00	4.18	2.76
Exams completed year 2	494	0	11	5.00	3.96	3.34
Exams completed year 3	494	0	15	5.00	4.68	4.11
Exams completed year 4	494	0	14	0.00	1.92	2.79
Exams completed year 5	494	0	11	0.00	0.51	1.44
Training Credits Completed year 1	494	0	78	36.00	33.49	22.14
Training Credits Completed year 2	494	0	81	34.50	30.10	25.18
Training Credits Completed year 3	494	0	99	36.00	31.71	27.36
Training Credits Completed year 4	494	0	99	0.00	13.30	19.36
Training Credits Completed year 5	494	0	81	0.00	3.50	10.13
Training Credits Left Year 1	494	0	54	18.00	20.51	22.14
Training Credits Left Year 2	494	0	63	28.50	32.90	25.18
Training Credits Left Year 3	494	0	63	27.00	31.29	27.36
Training Credits Left after 2 years	494	0	117	45.00	53.41	44.77
Training Credits Left after 3 years	494	0	180	69.00	84.70	69.56

3.2 Results

Multinomial Logistic Regression. The multinomial logistic regression model was initially estimated using all available data on students' academic careers during the first five years of study. This time frame was chosen to capture a sufficiently long trajectory, although voluntary dropout typically occurs during the second year.

Table 5. Effects of statistically significant covariates in the multinomial logistic regression model based on the first five years of the university career. Odds Ratios (OR) and 95% Confidence Intervals (CI) are reported ($p < 0.05$).

	status = 0		status = 2		status = 3	
	OR	95% CI	OR	95% CI	OR	95% CI
Degree Course*						
L025	0.925	0.127 – 6.753	0.073	0.006 – 0.858	0.066	0.005 – 0.801
L026	5.298	0.598 – 46.949	2.034	0.086 – 48.215	7.563	0.356 – 160.836
Credits Left Year 5	1.118	1.078 – 1.161	1.174	1.127 – 1.224	1.172	1.125 – 1.222

Reference categories: status=1, * L021
Hosmer and Lemeshow Goodness-of-Fit Test p=0-3797

Table 5 tells us that as the number of training credits left at the fifth year increases, the probability of being in status 0 (OR=1.118) and especially in status 2 (OR = 1.174) and 3 (OR = 1.172) also increases. Moreover, comparing the courses of study, course L021 is associated with a higher relative odds of failure than course L025. In fact, the odds ratios for status 2 (OR = 0.073, CI = 0.006-0.858) and status 3 (OR = 0.066, CI = 0.005–0.801) have values significantly less than 1. However, from Table 1 it is clear that most students in status 2 tend to drop out of their studies by their second year. Therefore, to evaluate whether early academic indicators could improve predictive performance, a second specification was estimated using academic data from only the first two years of study.

This approach allows the model (Table 6) to capture early signs of attrition and assess whether information from the initial years of university can enhance the prediction of different academic outcomes.

Table 6. Effects of statistically significant covariates in multinomial logistic regression considering among the other variables the first two years of the University career: Odds Ratio (OR) and Wald Confidence Interval (CI) are reported ($p < 0.05$).

	status = 0		status = 2		status = 3	
	OR	95% CI	OR	95% CI	OR	95% CI
Degree Course*						
L025	1.072	0.324 – 3.541	0.052	0.008 – 0.328	0.053	0.008 – 0.344
L026	1.015	0.204 – 5.059	0.194	0.023 – 1.678	0.786	0.112 – 5.513
Credit Left after 2 years	1.064	1.042 – 1.087	1.082	1.051 – 1.114	1.093	1.062 – 1.125
Exam Grade Year 2	0.996	0.906 – 1.093	0.846	0.762 – 0.939	0.884	0.798 – 0.978

Reference categories: status=1, * L021
Hosmer and Lemeshow Goodness-of-Fit Test p=0.9832

The model estimated using the first two years (Table 6) confirms persistent differences between the two degree programs,L025 and L021, particularly in the early identification of dropout risk. Holding other covariates constant, an increase in the number of credits remaining at the end of the first two years is associated with a higher probability of dropout. Conversely, a higher weighted average of grades in the second year is negatively associated with the likelihood of experiencing outcomes corresponding to status 2 or 3, suggesting a protective effect of strong academic performance during the early stages of the degree program. Nevertheless, the Hosmer and Lemeshow Goodness-of-Fit test states the superiority of the first model (Table 5), and the Brier Score on the validation data backs it up as well (0.26 against 0.34).

Competing Risks. The Fine–Gray competing risks model (Table 7) provides a dynamic representation of academic trajectories accounting for the presence of multiple mutually exclusive outcomes. Results indicate that higher values of the covariates are associated with an increased subdistribution hazard of graduation (status 1), in particular higher weighted average grades in the third year (HR = 1.055, CI = 1.021–1.091) and higher secondary school graduation marks (HR = 1.131, CI = 1.096–1.167).

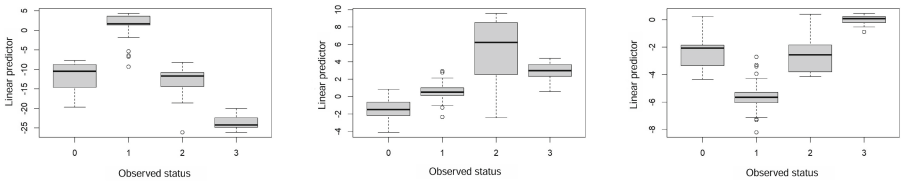
Table 7. Effect of covariates in Fine–Gray subdistributional hazard regression model: Subdistributional Hazard Ratio (HR) and 95% Confidence Interval (CI)

	status = 1		status = 2		status = 3	
	HR	95% CI	HR	95% CI	HR	95% CI
Degree Course*						
L025					0.609	0.259 – 1.436
L026					1.594	1.151 – 2.209
School Grad Mark	1.131	1.096 – 1.167	1.051	1.039 – 1.063	0.962	0.950 – 0.973
Time-by-Mark interact.	0.917	0.895 – 0.939	0.959	0.953 – 0.965	1.023	1.018 – 1.028
Mean Year 3	1.055	1.021 – 1.091				
Credits Compl. Year 1				0.974	0.956	0.991
Credits Compl. Year 3				0.959	0.934	0.986
Credits Compl. Year 4				0.936	0.891	0.982
Credits Left Year 2			1.082	1.065 – 1.101		
Credits Left Year 5	0.883	0.847 – 0.921				

Reference category: * L021

Conversely, it decreases as the number of remaining training credits in the fifth year increases (HR = 0.883, CI = 0.847–0.921), suggesting that delayed academic progress is associated with a lower likelihood of degree completion. For dropout (status 2), the subdistribution hazard increases with the number of credits sill to be completed in the second year (HR = 1.082, CI = 1.065–1.101) and with higher secondary school graduation marks (HR = 1.051, CI = 1.039–1.060). However,

the effect of secondary school performance decreases over time, as indicated by a subdistribution hazard ratio below one for the corresponding interaction term with time. Regarding inactivity-related withdrawal (status 3), students enrolled in the L026° program exhibit a higher risk compared to those in L021 (HR = 1.594, CI =1.151–2.209). In contrast, a higher accumulation of credits in the first, third, and especially fourth years is associated with a reduced risk of this outcome (HR<1), indicating a protective effect of sustained academic progress. Secondary school graduation marks also show a protective association (HR = 0.962, CI= 0.950–0.973), although this effect varies over time. The boxplot analysis of the linear predictors provides additional insight into model discrimination (Fig. 1). For statuses 1 (graduation) and 3 (inactivity), a clear separation between outcome groups is observed, indicating strong discriminative ability. In contrast, status 2 (dropout) shows partial overlap between categories, reflecting the intermediate and heterogeneous nature of this outcome.



(a) Graduation (b) Dropout (c) Inactivity

Fig. 1. Distribution of the estimated event-specific linear predictors across observed outcome categories.

Table 8. Time-dependent AUC values of the Fine–Gray competing risks model on the validation set.

Event	Time (years)							
	2	3	4	5	6	7	8	9
Graduation (Status 1)	–	–	1.000	1.000	1.000	1.000	1.000	0.997
Dropout (Status 2)	1.00	1.00	1	0.988	0.984	0.987	0.966	0.965
Inactivity (Status 3)	–	–	–	0.871	0.897	0.976	0.977	0.980

The predictive performance of the Fine–Gray competing risks model was assessed on the validation sample using time-dependent AUC and the distribution of linear predictors across observed outcomes. The results (Table 8) show a generally high discriminative ability across all event types, although with different patterns over time.

For graduation (status 1), AUC values are close to 1 from year 4 onwards, indicating an almost perfect ability of the model to distinguish between students who graduate and those who do not at later stages of the academic trajectory. This result is consistent with the structure of the data, as no graduation events are observed in the early years.
For dropout (status 2), the model exhibits very high discrimination at early time points (AUC = 1 at years 2 and 3), followed by a slight decrease over time, with values stabilizing around 0.96–0.99. This pattern suggests that early dropout is highly predictable, while later dropout becomes more heterogeneous and therefore harder to distinguish.
For inactivity-related withdrawal (status 3), AUC values are not available in the initial years due to the absence of events, but show a progressive increase from year 5 onwards, reaching values above 0.97 in later periods. This indicates an improvement in predictive accuracy over time, likely reflecting the cumulative effect of academic performance indicators. These results should be interpreted in light of the relatively small number of events at specific time points, particularly in the early follow-up period. Overall, they confirm the strong discriminative ability of the model, while also highlighting differences in predictability across event types and time horizons.

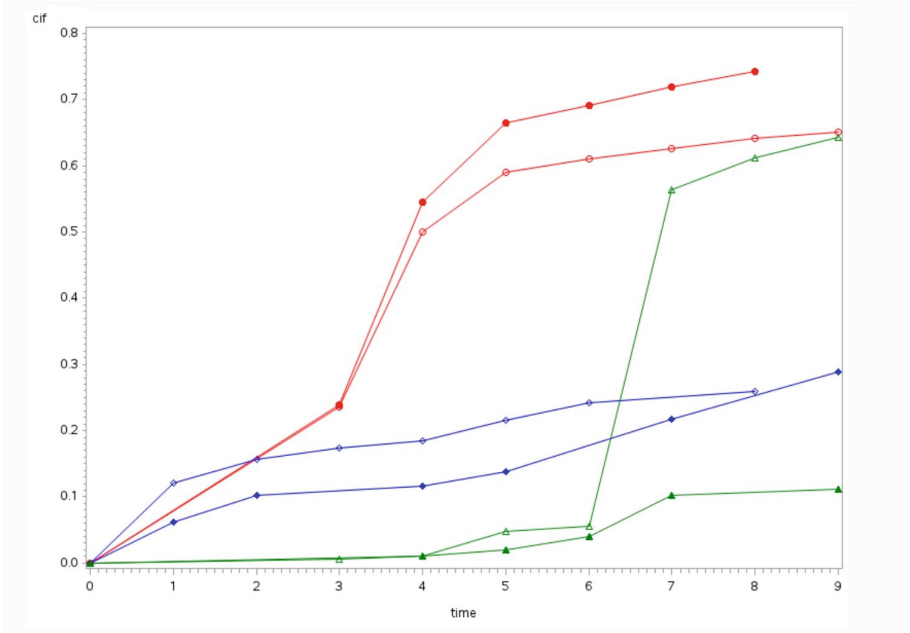


Fig. 2. Cumulative incidence functions (CIF) estimated for training (open symbols) and validation (filled symbols) samples under the Fine–Gray competing risks framework. Upper, intermediate, and lower curves correspond to Events 1 (red), 2 (blue), and 3 (green), respectively. (Color figure online)

Finally, the comparison of cumulative incidence functions between training and validation samples (Fig. 2) provides a qualitative assessment of model calibration over time. Good agreement is observed for graduation (status 1) and inactivity-related withdrawal (status 3) in the early follow-up period, while discrepancies emerge at later times. For dropout, a systematic difference between training and validation curves suggests lower predictive stability for this outcome.

Overall, these results confirm that academic performance indicators are the strongest determinants of student trajectories, while pre-university characteristics mainly act as initial conditions whose effect diminishes over time. The Fine–Gray model further highlights the importance of accounting for competing risks and temporal dynamics in the analysis of academic progression.

4 Conclusions

This study extends previous research by integrating pre-university characteristics with longitudinal information on students' academic trajectories, enabling a direct comparison between baseline attributes and time-varying performance indicators. The results indicate that pre-university characteristics (e.g., secondary school graduation marks and school type) are informative at entry but their relative contribution to prediction decreases as students progress through their academic careers. This pattern is consistent with prior literature [2,20] and supports the interpretation of these variables as initial conditions, whereas university performance reflects a dynamic process of academic integration [19,20].

Empirically, academic performance indicators—particularly accumulated and remaining credits—emerge as the strongest predictors of both graduation and dropout. Once in-course performance is accounted for, the explanatory power of pre-entry characteristics is substantially reduced. Differences across degree programs are also evident: L021 is associated with a higher risk of failure and dropout compared to L025, while L026 shows a higher likelihood of inactivity-related withdrawal. These findings highlight the importance of institutional and curricular structures in shaping student trajectories.

From a methodological perspective, the comparison between multinomial logistic regression and the Fine–Gray competing risks model reveals complementary advantages [10,16]. The multinomial framework provides a global view of academic outcomes and achieves strong predictive performance when full academic trajectories are used, as reflected by validation metrics such as the Brier Score. However, it offers limited insight into early-stage dropout dynamics. In contrast, the Fine–Gray model explicitly accounts for the temporal structure of the data and the presence of competing risks, providing a more appropriate framework for analysing student progression over time. Overall, there is a clear trade-off between predictive accuracy and temporal interpretability: full-trajectory models yield higher accuracy, whereas early-stage models are more suitable for timely identification of at-risk students.

The competing risks analysis further supports these findings. Higher academic performance increases the probability of graduation, whereas a greater accumulation of remaining credits is associated with delayed completion or dropout. Pre-university performance has a positive effect on academic success and reduces the risk of dropout and inactivity; however, its influence diminishes over time. In particular, strong early academic performance acts as a protective factor against attrition [8,12], underscoring the importance of early academic integration.

From a policy perspective, the results suggest that interventions based exclusively on pre-university profiles are insufficient for supporting long-term student success. Continuous monitoring of early academic performance provides a more reliable basis for identifying at-risk students and designing targeted interventions such as tutoring, mentoring, and personalised academic support. The observed heterogeneity across degree programmes further indicates that support strategies should be tailored to specific curricular contexts rather than applied uniformly.

The comparison of cumulative incidence functions between training and validation samples provides evidence of good model calibration in the early follow-up period, while some discrepancies emerge in later years, particularly for dropout-related outcomes. This graphical assessment complements the regression results by supporting the evaluation of model stability and generalisability. Consistently, time-dependent AUC values indicate strong discriminative performance of the Fine–Gray model, particularly for graduation and dropout at later time points, with gradual improvement for inactivity-related outcomes.

Finally, the distribution of linear predictors shows clear separation between outcome groups for graduation and inactivity-related withdrawal, while greater overlap is observed between dropout and inactivity. This suggests that dropout trajectories share characteristics with multiple academic pathways, making them more difficult to distinguish. Taken together, AUC measures [4], predictor distributions, and cumulative incidence comparisons provide a consistent assessment of model performance, indicating strong overall discrimination but also highlighting areas of reduced predictive clarity.

Several limitations should be acknowledged. The analysis relies on administrative data from a single institution and does not include potentially relevant variables such as socio-economic background, motivation, or student engagement. Future research should extend this framework to multi-institutional settings and incorporate richer individual-level information to improve predictive performance and generalisability

References

1. Bean, J.P.: Dropouts and turnover: the synthesis and test of a causal model of student attrition. Res. High. Educ. **12**(2), 155–187 (1980)
2. Bean, J.P.: Interaction effects based on class level in an explanatory model of college student dropout syndrome. Am. Educ. Res. J. **22**(1), 35–64 (1985)

3. Bettinger, E.P., Long, B.T., Oreopoulos, P., Sanbonmatsu, L.: The role of application assistance and information in college decisions: results from the H & R Block FAFSA experiment. Quart. J. Econ. **127**(3), 1205–1242 (2012)
4. Blanche, P., Dartigues, J.F., Jacqmin-Gadda, H.: Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat. Med. **32**(30), 5381–5397 (2013)
5. Blanche, P., Dartigues, J.F., Jacqmin-Gadda, H.: timeROC: Time-Dependent ROC Curve Estimation from Censored Survival Data (2024). https://CRAN.R-project.org/package=timeROC
6. Bound, J., Lovenheim, M.F., Turner, S.: Why have college completion rates declined? an analysis of changing student preparation and collegiate resources. Am. Econ. J. Appl. Econ. **2**(3), 129–157 (2010)
7. Buenaño, E., Beletanga, M.J., Mancheno, M.: What factors are relevant to understanding dropout? analysis at a co-financed university in ecuador and policy implications, using survival cox models. J. Latinos Educ. **23**(4), 1400–1415 (2024). https://doi.org/10.1080/15348431.2023.2271570
8. DesJardins, S.L., Ahlburg, D.A., McCall, B.P.: A temporal investigation of factors related to timely degree completion. J. Higher Educ. **73**(5), 555–581 (2002)
9. Fierro Saltos, W.R., Fierro Saltos, F.E., Elizabeth Alexandra, V.S., Rivera Guzmán, E.F.: Leveraging artificial intelligence for sustainable tutoring and dropout prevention in higher education: a scoping review on digital transformation. Information **16**(9), 819 (2025)
10. Fine, J.P., Gray, R.J.: A proportional hazards model for the subdistribution of a competing risk. J. Am. Stat. Assoc. **94**(446), 496–509 (1999)
11. Heisey, D.M., Patterson, B.R.: A review of methods to estimate cause-specific mortality in presence of competing risks. J. Wildl. Manag. **70**(6), 1544–1555 (2006)
12. Herzog, S.: Measuring determinants of student return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshmen. J. College Student Reten. Res. Theory Pract. **6**(1), 57–74 (2005)
13. Hosmer, D.W., Lemeshow, S.: Goodness of fit tests for the multiple logistic regression model. Commun. Stat. Theory Methods **9**(10), 1043–1069 (1980)
14. Leow, T., Li, W.W., Miller, D.J., McDermott, B.: Prevalence of university non-continuation and mental health conditions, and effect of mental health conditions on non-continuation: a systematic review and meta-analysis. J. Ment. Heal. **34**(2), 222–237 (2025). https://doi.org/10.1080/09638237.2024.2332812, pMID: 38588717
15. Pusztai, G., Fényes, H., Kovács, K.: Factors influencing the chance of dropout or being at risk of dropout in higher education. Educ. Sci. **12**(11) (2022). https://doi.org/10.3390/educsci12110804, https://www.mdpi.com/2227-7102/12/11/804
16. Putter, H., Schumacher, M., van Houwelingen, H.C.: On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: the fine-gray model revisited. Biom. J. **62**(3), 790–807 (2020)
17. SAS: SAS/STAT® 9.22 User's Guide. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect004.htm. Accessed 27 Nov 2025
18. SAS: SAS/STAT® 9.22 User's Guide. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#phreg_toc.htm. Accessed 26 Jan 2026
19. Tinto, V.: Dropout from higher education: a theoretical synthesis of recent research. Rev. Educ. Res. **45**(1), 89–125 (1975). https://doi.org/10.3102/00346543045001089

20. Tinto, V.: Leaving College: Rethinking the Causes and Cures of Student Attrition. University of Chicago Press, Chicago (1993)
21. Vaarma, M., Li, H.: Predicting student dropouts with machine learning: an empirical study in finnish higher education. Technol. Soc. **76**, 102474 (2024). https://doi.org/10.1016/j.techsoc.2024.102474, https://www.sciencedirect.com/science/article/pii/S0160791X24000228

Dual-Branch Hidden Markov Model for Rolling Bearing Degradation Analysis

Indrawata Wardhana[1,2(✉)], Amal Gouiaa-Mtibaa[3], Pascal Vrignat[3], and Frédéric Kratz[1]

[1] INSA Centre Val de Loire, PRISME Laboratory, 18000 Bourges, France
Indrawata.wardhana@insa-cvl.fr
[2] UIN Sulthan Thaha Saifuddin Jambi, 36361 Muaro Jambi, Indonesia
[3] Orléans University, PRISME Laboratory, 45000 Orléans, France

Abstract. Accurate identification of bearing faults remains a challenging task in rotating machinery monitoring due to noise contamination and pseudo-cyclostationary behavior, which obscure fault-related features, particularly in early degradation stages. To address these challenges, this paper proposes a Dual-Branch Hidden Markov Model (DB-HMM) for bearing fault diagnosis and degradation analysis. The proposed framework integrates two complementary feature branches to capture both impulsive characteristics and harmonic structures associated with characteristic fault frequencies. A dual-state formulation is introduced to enable simultaneous modeling of multiple fault components, specifically Ball Pass Frequency Outer (BPFO) and Ball Pass Frequency Inner (BPFI). By combining dual-branch feature representation with probabilistic sequential modeling, the proposed method effectively captures both temporal degradation evolution and complex signal structures. Experimental results demonstrate that the DB-HMM improves fault discrimination and enables more reliable and informative identification of degradation stages under noisy and non-stationary conditions.

Keywords: Hidden Markov Models · Envelope Signal · Bearing · Degradation Estimation

1 Introduction

Reliable detection of defects in rolling bearings remains a challenging task in the monitoring of rotating machinery. In practical industrial environments, vibration signals generated by faulty bearings are often contaminated by noise, structural vibrations, and operational variations, which may obscure fault-related signatures. Undetected defects may progressively propagate and eventually lead to severe machine failures or unexpected system shutdowns. Consequently, accurate fault diagnosis and degradation monitoring are essential for predictive maintenance and system reliability [1].

When a localized defect occurs on bearing components such as the inner race, outer race, or rolling elements, repeated interactions between rolling elements and the damaged surface generate impact events during rotation. These impacts excite the resonance

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 20–36, 2027.
https://doi.org/10.1007/978-3-032-30521-3_2

modes of the surrounding mechanical structure and produce vibration signals containing useful diagnostic information about the fault. Due to the periodic motion of rolling elements relative to the defect location, the resulting vibration signals often exhibit cyclostationary characteristics. This periodic behavior produces characteristic fault frequencies (FCFs), which are commonly used as indicators for detecting and diagnosing bearing faults [2].

However, in practical operating conditions, bearing vibration signals rarely exhibit ideal cyclostationary behavior. Instead, they often demonstrate quasi-cyclostationary or pseudo-cyclostationary characteristics as shown in Fig. 1 (a). These irregularities are mainly caused by random slips between rolling elements and raceways, as well as differences in rolling speed between loaded and unloaded zones. As a result, the impact sequence generated by defective bearings may not occur in a perfectly periodic manner. This deviation from ideal cyclostationarity complicates the detection of fault-related components in the vibration signal. The problem becomes even more challenging during the early stage of degradation, where the energy of fault-related frequencies is typically very weak and can easily be masked by background noise and structural vibrations.

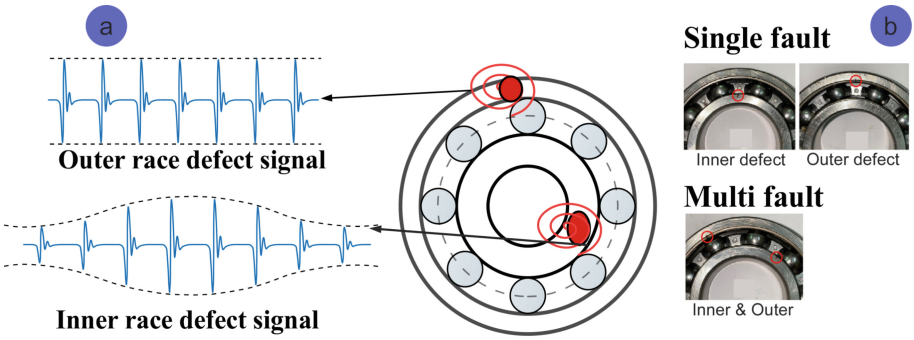


Fig. 1. (a) Pseudo-cyclostationary of bearing faults (b) Location of bearing defects [3].

In addition to pseudo-cyclostationary behavior, another difficulty arises from the complexity of fault patterns in practical bearing systems. A single localized defect usually generates periodic impacts at a specific characteristic frequency associated with the defective component. As the defect severity increases, the amplitude of the vibration response also increases. However, in many practical cases, bearings may experience multi-point defects or distributed damage on the raceway surface as illustrated in Fig. 1 (b). Such conditions can produce overlapping spectral components and nonlinear vibration responses. The interaction between multiple defects may lead to complex vibration signatures that are difficult to interpret using conventional diagnostic approaches.

Despite these challenges, characteristic fault frequencies remain an important indicator for bearing fault diagnosis. When defects evolve, harmonic components of the FCF gradually appear in the vibration spectrum [4]. These harmonic components reflect the repetitive interaction between rolling elements and defect locations. In the early stage of degradation, the harmonic components are typically weak and limited in number. As the defect grows, both the amplitude and the number of observable harmonics tend

to increase. Severe defects often produce strong higher-order harmonics, which indicate advanced damage conditions [5]. Therefore, the harmonic structure associated with FCFs can provide valuable information about the progression of bearing degradation.

Many existing approaches attempt to quantify bearing degradation using a Health index (HI) [6]. A HI is usually constructed from vibration features and is designed to represent the overall condition of the bearing. While HI-based approaches can provide a general indication of system deterioration, they often lack the ability to identify the specific location of the defect within the bearing components [7]. In other words, the health index may indicate that degradation is occurring, but it does not explicitly distinguish whether the fault originates from the inner race, outer race, or rolling elements [8]. Furthermore, many HI-based degradation models assume a monotonic degradation trend, where the health indicator gradually decreases from a healthy state toward failure [9]. In real industrial environments, however, degradation processes may not follow such an ideal monotonic behavior due to noise, load variations, and operational uncertainties.

To improve fault detection performance, various signal processing and machine learning methods have been proposed for bearing fault diagnosis [10, 11]. Traditional signal processing techniques include envelope analysis, spectral kurtosis, and time–frequency analysis, which aim to enhance impulsive components generated by localized defects. More recently, machine learning approaches such as support vector machines [12], neural networks [13], and deep learning models [14] have been applied to classify bearing fault conditions. Although these approaches can achieve high classification accuracy, many of them require large labeled datasets and may not explicitly model the temporal evolution of degradation processes.

Sequential probabilistic models provide an effective framework for analyzing time-dependent degradation processes. Among these methods, Hidden Markov Models (HMMs) are particularly suitable for modeling systems whose underlying states cannot be directly observed but can be inferred from measurable signals. In the context of bearing diagnostics, the hidden states can represent different degradation stages, while the observed vibration features provide indirect evidence of these states. Through probabilistic state transitions, HMMs can effectively capture the temporal evolution of degradation processes without relying on strictly monotonic HI.

However, conventional HMM-based approaches often rely on a single feature representation extracted from vibration signals [15]. In practice, bearing faults manifest through multiple signal characteristics, including impulsive responses, harmonic structures related to FCFs, and energy variations in specific frequency bands [16]. Relying on a single feature representation may therefore limit the ability of the model to capture the complex dynamics of bearing vibration signals, particularly under pseudo-cyclostationary conditions and noisy environments.

To address these limitations, this study proposes a DB-HMM framework for bearing fault diagnosis and degradation analysis. The proposed method integrates two complementary feature branches that capture impulsive behavior and harmonic characteristics associated with fault characteristic frequencies. In this work, the analysis focuses primarily on BPFO and BPFI components, whose spectral peaks are generally clearer and more reliable for degradation monitoring compared to Ball Spin Frequency, which

often exhibits weaker and less stable spectral signatures. By combining dual-branch feature representations with a probabilistic sequential modeling framework, the proposed method can capture both the temporal evolution of degradation and the complex structure of vibration signals.

The main contributions of this work are summarized as follows:

First, a novel DB-HMM framework is proposed to model the degradation process of rolling bearings using dual-branch feature representations.
Second, the proposed approach leverages harmonic characteristics of BPFO and BPFI components to represent the evolution of bearing degradation stages.
Third, the framework enhances robustness against pseudo-cyclostationary behavior, noise interference, and complex fault interactions.
Finally, experimental results demonstrate that the proposed method achieves improved performance for bearing fault diagnosis and degradation stage identification.

The remainder of this paper is organized as follows. Section 2 introduces the characteristics of bearing vibration signals and the background of pseudo-cyclostationary behavior. Section 3 presents the proposed DB-HMM framework and the feature extraction methodology. Section 4 describes the experimental setup and datasets used for validation. Section 5 presents the experimental results and discusses the performance of the proposed method. Finally, Sect. 6 concludes the paper and outlines directions for future work.

2 Methodology

The overall framework of the proposed DB-HMM is illustrated in Fig. 2. The method integrates two features of extraction branches, namely the envelope-based branch and the harmonic-based branch, which are combined through probabilistic modeling using HMM.

A synthetic vibration dataset is generated to simulate bearing degradation under controlled conditions, including both single-fault and multi-fault scenarios. The simulation equations are derived from industry-representative bearing vibration models and operational characteristics to ensure realistic degradation behavior. This approach enables flexible generation of both single-fault and multiple-fault degradation patterns in the desired format, without relying on publicly available benchmark datasets such as the Case Western Reserve University or Intelligent Maintenance Systems Center datasets. The main parameters used for signal generation, including system configuration and bearing geometry, are summarized in Table 1 [17].

Two degradation scenarios are considered: single-fault and multi-fault conditions. For single-fault scenarios, 20 samples are generated for each degradation state. For multi-fault scenarios, 20 samples are generated for each fault combination.

Single fault:

BPFO degradation,
BPFI degradation.

Multi fault:

Table 1. System configuration and bearing parameters

Parameter	Value	Parameter	Value
Sampling frequency F_S	25600 Hz	Number of balls n	8
Signal duration T	1s	Ball diameter d	7.94
Number of samples N	$F_S \bullet T$	Pitch diameter D	34.55
Rotational speed	2000 RPM	Contact angle θ	0
Rotational frequency f_S	RPM /60	Noise level σ	0.3

BPFO: Severe, BPFI: Early,
BPFO: Medium, BPFI: Failure,
BPFO: Failure, BPFI: Medium.

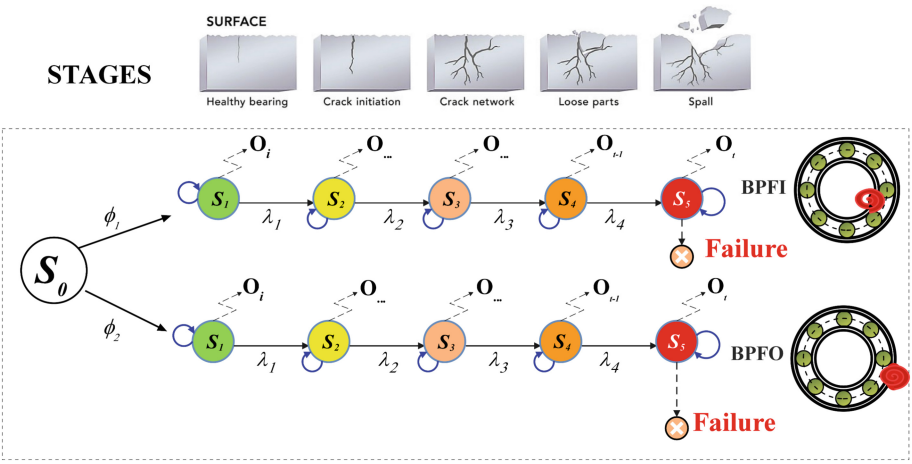


Fig. 2. Dual Branch HMM method

2.1 Bearing Fault Signal Modeling

The vibration signal of a rolling bearing is modeled as [17]:

$$x(t) = b(t) + r(t) + h(t) + n(t) \tag{1}$$

where $b(t)$ is the periodic fault impulse, $r(t)$ represents random impulses, $h(t)$ denotes harmonic components, and $n(t) \sim \mathcal{N}(0, \sigma^2)$ is additive Gaussian noise.

2.2 Fault Impulse Modeling

The fault-induced impulsive component is defined as:

$$b(t) = \sum_i A \bullet e^{-\beta_1 * \left(t - \frac{i}{f_m}\right)} \bullet sin\left(2\pi f_r \left(t - \frac{i}{f_m}\right)\right) \tag{2}$$

Two fault types are considered: Outer race fault $A_i \sim \mathcal{U}(0.9,1)$ and Inner race fault $A_i = \frac{1-sin(2\pi f_r t)}{2}$, where f_m is the fault characteristic frequency and f_r is the rotational frequency.

2.3 Harmonic Components

The harmonic component is defined as:

$$h(t) = \sum_{k=1}^{K} Psin(2\pi h_k t + \theta_k) \tag{3}$$

2.4 Envelope Signal and Generalized Envelope Spectrum

The analytic signal is defined using the Hilbert transform:

$$\alpha(t) = x(t) + j\mathcal{H}(x(t)) \tag{4}$$

where $\mathcal{H}$ denotes the Hilbert transform. The conventional envelope signal is:

$$E(t) = \alpha(t) = \sqrt{x^2(t) + \mathcal{H}^2\{x(t)\}} \tag{5}$$

The generalized envelope is defined using a parameter υ [18]:

$$GE_\upsilon(t) = \begin{cases} log(\lceil \alpha(t) \rceil), & \nu = 0 \\ \lceil \alpha(t) \rceil^\upsilon, & \nu > 0 \end{cases} \tag{6}$$

The corresponding Generalized Envelope Spectrum (GES) is obtained via Fourier transform:

$$GES_\upsilon(f) = \mathcal{F}\{GE_\upsilon(t)\} \tag{7}$$

where $\mathcal{F}(\bullet)$ denotes the Fourier transform. The parameter ν controls feature enhancement, where $\nu > 1$ improves robustness to noise, while $\nu < 1$ enhances impulsive components.

2.5 Feature Extraction

The degradation characteristics are represented using logarithmic and ratio-based harmonic features:

$$F_1 = log_{10}(H_1)\ F_2 = log(H_2 + H_3 + H_4) \tag{8}$$

The feature vector is given by $X = [F_1, F_2]$. For clustering, a reduced feature space is adopted $X_c = [F_1, F_5]$.

2.6 Gaussian Mixture Model

The probability density of the feature vector is modeled using a mixture of Gaussian distributions [19]:

$$p(x) = \sum_{i=1}^{k} w_i \mathcal{N}(x|\mu_i, \Sigma_i) \tag{9}$$

where w_i, μ_i, and Σ_i denote the mixture weight, mean vector, and covariance matrix of the i-th Gaussian component. Each sample is assigned to a discrete observation symbol:

$$o_t = \underset{i}{argmax} P(i|x_t) \tag{10}$$

This process converts continuous features into discrete states suitable for HMM modeling. The observation sequence is defined as:

$$o = \{o_1, o_2, \ldots, o_T\} \tag{11}$$

2.7 Dual-Branch Hidden Markov Model

To model the temporal evolution of bearing degradation, a HMM is employed. The HMM is defined by three parameters [20]:

$$\lambda = (A, B, \pi) \tag{12}$$

where A: state transition probability matrix, B: observation probability matrix, π: initial state distribution

State Definition

$$s_1, s_2, s_3, s_4, s_5 = \{Healthy, Early, Medium, Severe, Failure\} \tag{13}$$

These states model the progressive evolution of bearing damage from a healthy condition to failure. For two Branch states are defined as:

$$s_t = \left(s_t^{(1)}, s_t^{(2)}\right) \tag{14}$$

where $s_t^{(1)}$ and $s_t^{(2)}$ represent the degradation states of Branch 1 and Branch 2, respectively.

State Transition Model

The transition probability matrix is defined as:

$$P\left(s_{t+1}|s_t\right) = P\left(s_{t+1}^{(1)}\middle|s_t^{(1)}\right) \cdot P\left(s_{t+1}^{(2)}\middle|s_t^{(2)}\right)\Big)\Big)\Big) \tag{15}$$

Each transition matrix follows a left-to-right structure:

$$a_{ij}^{(k)} = 0 for j < i, k \in \{1,2\} \tag{16}$$

This constraint ensures that the degradation process is irreversible.

Dual-Branch Observation Modeling

Unlike conventional HMM approaches that rely on a single feature representation, the proposed method employs two complementary observation Branches: Branch 1 (BPFO) and Branch 2 (BPFI). Each Branch generates an observation sequence:

$$o^{(1)} = \left\{o_1^{(1)}, \ldots, o_T^{(1)}\right\}, o^{(2)} = \left\{o_1^{(2)}, \ldots, o_T^{(2)}\right\} \tag{17}$$

The observation probability for each state is modeled as the joint likelihood of both Branches:

$$P(o_t|s_t) = P\left(o_t^{(1)}|s_t\right) \cdot P\left(o_t^{(2)}|s_t\right) \tag{18}$$

This formulation assumes conditional independence between the two branches while enabling probabilistic fusion of complementary information. Although both branches are derived from the same vibration signal. After conditioning on the hidden health state, both branches are assumed to provide complementary diagnostic information with reduced dependency.

From a signal-processing perspective, vibration signals contain multiple overlapping fault components in both temporal and frequency domains. The separation of BPFO and BPFI components through dedicated feature extraction reduces the correlation between observation branches and supports effective probabilistic fusion. Therefore, the conditional independence assumption is adopted as a practical simplification for the proposed HMM framework.

Observation Probability Matrix.

For each branch, the observation probability matrix is defined as:

$$B^{(k)} = \left[b_j^{(k)}(m)\right], b_j^{(k)}(m) = P\left(o_t^{(k)} = m|s_t = j\right) \tag{19}$$

where $k \in \{1,2\}$ denotes the Branch index and m represents the discrete observation symbol obtained from GMM clustering. The final observation likelihood is obtained by combining both matrices:

$$b_j(m) = b_j^{(1)}(m) \cdot b_j^{(2)}(m) \tag{20}$$

Initial State Distribution and Prior Modeling

The initial state distribution is defined as:

$$\pi = \pi^{(1)} \cdot \pi^{(2)} \tag{21}$$

For the dual-branch formulation, the prior probability of observations is expressed as:

$$\pi^{(1)} = [1,0,\ 0,0,\ 0],\ \pi^{(2)} = [1,0,\ 0,0,\ 0] \tag{22}$$

This reflects the parallel contribution of both feature branches at the initial stage.

State Inference

Given the trained model, the most probable hidden state sequence is obtained using the Viterbi algorithm:

$$s^* = \underset{s}{argmax} P(s|o, \lambda) \tag{23}$$

where $s = \{s_1, s_2, \ldots, s_T\}$ represents the hidden degradation states. This enables tracking of degradation progression over time.

Interpretation

This formulation allows simultaneous tracking of degradation states, for example:

$$(Normal, Normal) \rightarrow (Medium, Early) \rightarrow (Failure, Medium) \tag{24}$$

3 Results and Discussion

This section presents a comprehensive evaluation of the proposed Dual-Branch Hidden Markov Model (DB-HMM) for bearing degradation analysis. The analysis is conducted from multiple perspectives, including time-domain signal characteristics, spectral behavior, feature representation, multi-fault tracking capability, transition dynamics, and classification performance.

3.1 Time-Domain Signal Characteristics

The time-domain vibration signals provide direct insight into the evolution of fault-induced impulses. Figure 3 shows the BPFO signals under different degradation stages. In the early stage, the signal exhibits weak and sparse impulsive components, which are often masked by noise. As degradation progresses, the impulses become more frequent and exhibit higher amplitudes, indicating stronger interactions between rolling elements and defect surfaces. In the severe and failure stages, impulsive behavior becomes highly periodic and dominant, reflecting significant structural damage.

In contrast, Fig. 4 presents the vibration signals for BPFI. Due to the rotating nature of inner race defects, the signal exhibits amplitude modulation effects, resulting in more irregular and less stable patterns. This behavior makes early fault detection more challenging compared to BPFO.

Figure 5 illustrates vibration signals under combined fault conditions. In the first case (Fig. 5 (a)), the signal is dominated by strong and highly periodic impulsive components. This behavior is mainly associated with the severe BPFO fault, which produces consistent high amplitude impacts due to repeated interactions between rolling elements and the outer race defect. The presence of the BPFI medium fault introduces additional modulation, although its contribution is less dominant.

In contrast, the second case (Fig. 5 (b)) exhibits more irregular and fluctuating impulsive patterns. The BPFI failure condition introduces strong amplitude modulation due to the rotating nature of the inner race defect, resulting in a less stable signal structure. Meanwhile, the BPFO medium fault contributes to weaker and less consistent impacts. These results highlight that different fault components contribute differently to the overall vibration signal depending on their severity levels. Outer race defects tend to produce

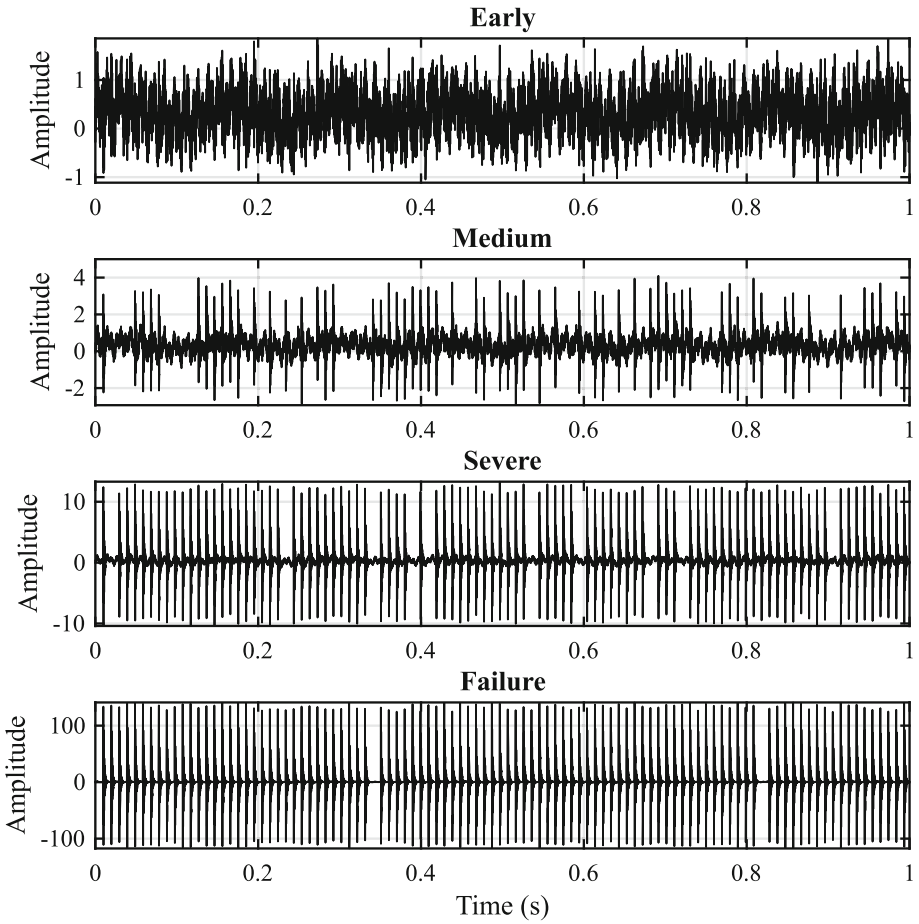


Fig. 3. Time-domain vibration signals under different fault severity levels (Early–Failure) for BPFO

more stable and periodic impulsive signatures, whereas inner race defects generate more complex and modulated patterns. This difference confirms the necessity of a dual-branch modeling approach. By separating the contributions of different fault components, the proposed DB-HMM can better capture heterogeneous degradation behaviors that cannot be represented using a single-state model.

To enhance fault-related features, the GES is applied. Figure 6 shows that the harmonic structure associated with BPFO becomes increasingly pronounced as degradation progresses. In the early stage, only low-order harmonics were visible with weak amplitudes. As the fault evolves, higher-order harmonics emerge and their amplitudes increase significantly, indicating stronger periodic impacts.

For BPFI Fig. 7, the harmonic structure is less stable due to pseudo-cyclostationary behavior caused by varying contact conditions. Despite this complexity, GES remains

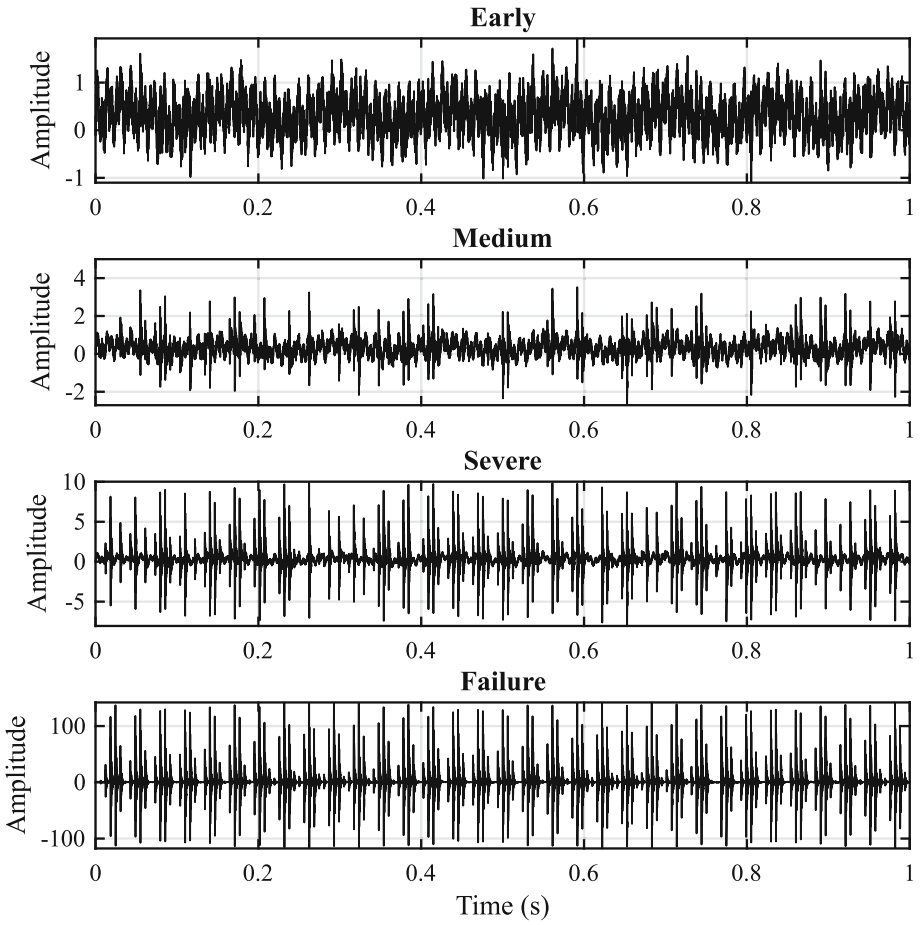


Fig. 4. Time-domain vibration signals under different fault severity levels (Early–Failure) for BPFI.

effective in highlighting fault-related components, demonstrating its robustness under non-ideal operating conditions.

The extracted harmonic features are transformed into a compact representation and clustered using a GMM, as illustrated in Fig. 8. This step converts continuous feature space into discrete observation symbols suitable for HMM modeling. As shown in Fig. 6(a), the clustering results for BPFO exhibit a clear and well-structured separation between degradation stages. The clusters are distributed along a monotonic progression, where samples corresponding to healthy conditions are concentrated in the lower feature region, while severe and failure states occupy higher feature values. The Gaussian components form compact and well-separated ellipsoidal regions, indicating low intra-cluster variance and strong discriminative capability. This behavior reflects the relatively stable and periodic impulsive characteristics of BPFO faults, which result in consistent harmonic patterns in the GES domain.

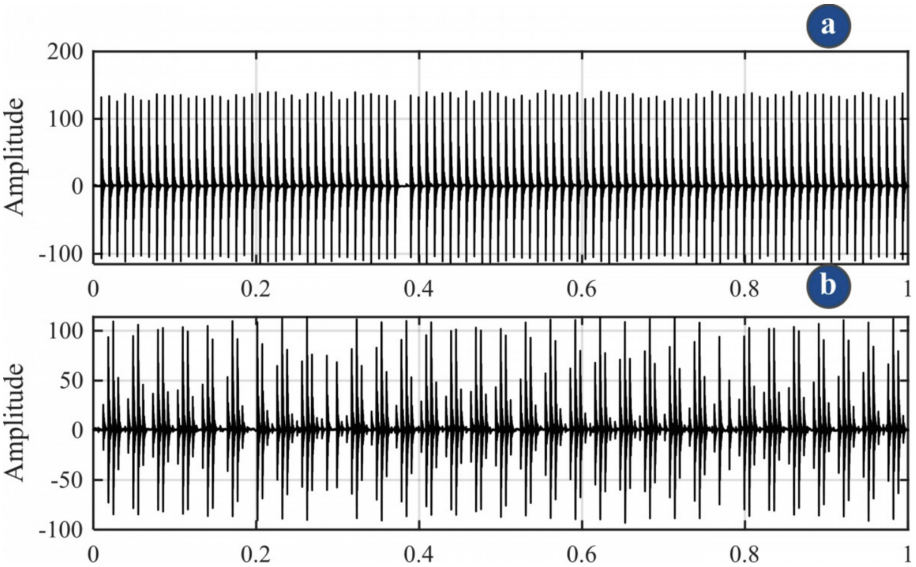


Fig. 5. Time-domain vibration signals (a) BPFO: Medium, BPFI: Failure (b) BPFO: Failure, BPFI: Medium

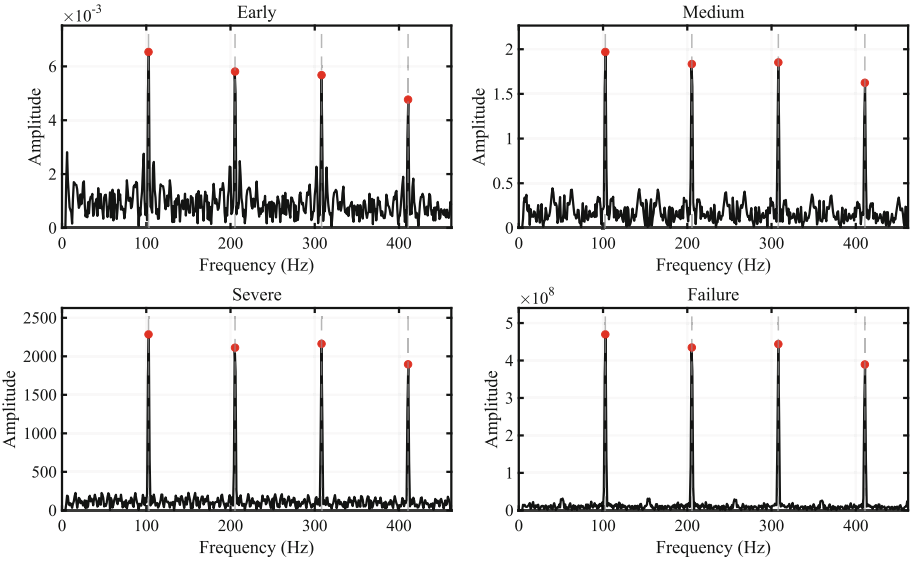


Fig. 6. GES of BPFO under different degradation stages

In contrast, Fig. 6(b) shows the clustering results for BPFI, which exhibit a more scattered and overlapping distribution. Although the overall progression from healthy to faulty states is still observable, the cluster boundaries are less distinct compared to BPFO. This behavior is attributed to the rotating nature of inner race defects, which

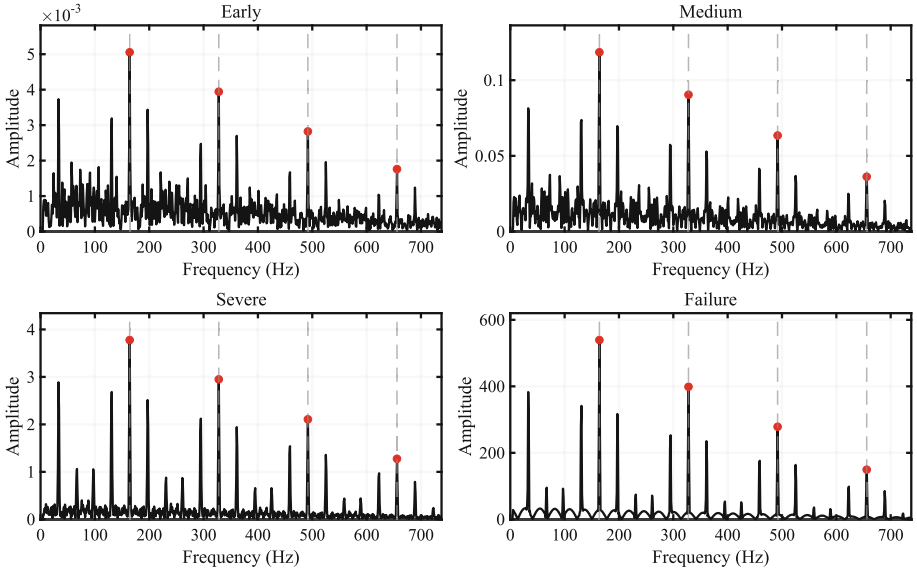


Fig. 7. GES of BPFI under different degradation stages

introduces amplitude modulation and pseudo-cyclostationary effects. As a result, the extracted features show higher variability, leading to increased overlap between adjacent degradation states.

Despite the increased variability in BPFI, the GMM is still able to capture the general trend of degradation and provide meaningful discretization. The ordering of clusters based on feature magnitude ensures a consistent mapping between observation symbols and degradation levels, which is essential for constructing reliable observation sequences in the DB-HMM framework.

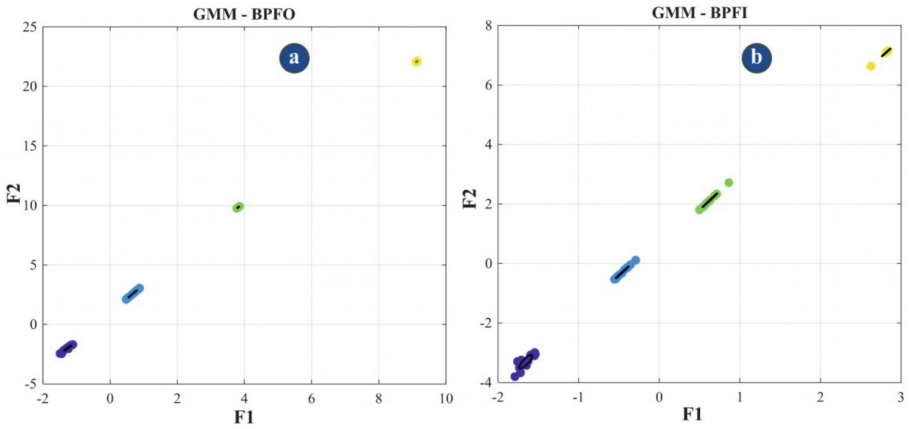


Fig. 8. GMM clustering results in the feature space: (a) BPFO and (b) BPFI.

The dataset is divided into training and testing sets using a 30:70 ratio. The training data are used to estimate the DB-HMM parameters, while the testing data are used for evaluation. The learned transition matrices provide important insight into degradation dynamics. Both matrices exhibit strong diagonal dominance, indicating that each degradation state persists over time before transitioning to the next stage. This behavior aligns with the physical nature of bearing degradation, which evolves gradually rather than abruptly. A representative transition matrix for Branch 1 and the transition matrix for Branch 2 are:

$$A^{(1)} = \begin{bmatrix} 0.992 & 0.008 & 0 & 0 & 0 \\ 0 & 0.933 & 0.067 & 0 & 0 \\ 0 & 0 & 0.929 & 0.071 & 0 \\ 0 & 0 & 0 & 0.929 & 0.071 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A^{(2)} = \begin{bmatrix} 0.991 & 0.009 & 0 & 0 & 0 \\ 0 & 0.920 & 0.080 & 0 & 0 \\ 0 & 0 & 0.915 & 0.085 & 0 \\ 0 & 0 & 0 & 0.905 & 0.095 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Furthermore, differences between $A^{(1)}$ and $A^{(2)}$ suggest that the degradation rates are not identical. Branch 2 shows higher self-transition probabilities in early states, indicating a slower and more gradual evolution compared to Branch 1. This confirms that the learned transition matrices are not arbitrary but reflect physically meaningful degradation behavior.

The classification performance is evaluated using confusion matrices, as presented in Fig. 9 for BPFO, BPFI, and multi-defect conditions. In both single-fault cases (Fig. 9 (a) and Fig. 9 (b)), the results show a perfect diagonal structure, indicating that all samples are correctly classified into their respective degradation states. This strong diagonal dominance confirms that the proposed feature representation effectively separates the degradation levels in the feature space. In particular, the distinction between Medium, Severe, and Failure states is highly consistent, reflecting the increasing strength of impulsive and harmonic components as the fault progresses.

For the multi-defect scenario (Fig. 9 (c)), the results also demonstrate accurate classification of combined fault states. The model successfully identifies different degradation levels between BPFO and BPFI simultaneously, as shown by the correct mapping of state combinations. This indicates that the proposed dual-branch framework can capture heterogeneous degradation behaviors across multiple fault components. Despite the inherent difficulty in detecting early-stage faults due to weaker signal characteristics, the proposed approach maintains high classification performance across all conditions. This highlights the benefit of combining complementary feature domains and probabilistic modeling. Overall, the results achieve an accuracy of 100% for all evaluated cases. It should be noted that the high accuracy is partly influenced by using synthetically generated data under controlled conditions, which ensures clear separation between degradation states.

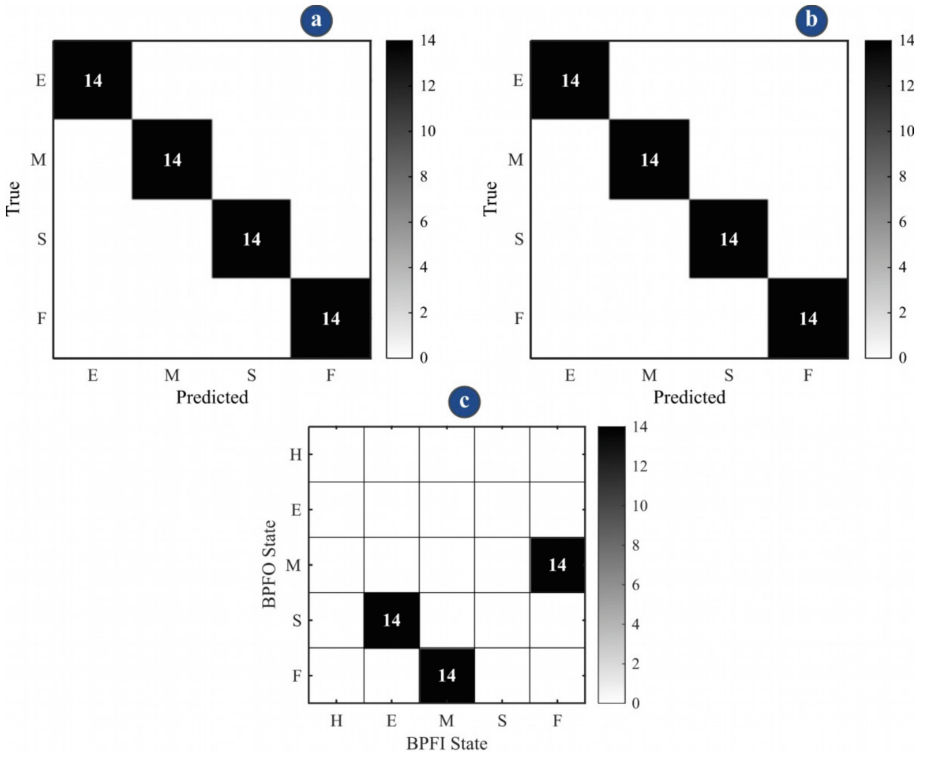


Fig. 9. Confusion matrices for bearing fault degradation classification: (a) BPFO, (b) BPFI, and (c) multi-defect BPFO–BPFI. The labels denote degradation states: Early (E), Medium (M), Severe (S), and Failure (F).

4 Conclusion

This paper proposed a DB-HMM for bearing degradation analysis. This method integrates two complementary feature branches and enables simultaneous tracking of multiple fault types through a dual-state formulation. Experimental results show that the proposed approach effectively captures degradation progression and improves fault discrimination under noisy and pseudo-cyclostationary conditions. Future work will focus on extending the framework to more complex multi-fault scenarios and real industrial applications.

References

1. Mohanty, S., Gupta, K.K., Raju, K.S.: Hurst based vibro-acoustic feature extraction of bearing using EMD and VMD. Meas. J. Int. Meas. Confed. **117**, 200–220 (2018)
2. Sacerdoti, D., Strozzi, M., Secchi, C.: A comparison of signal analysis techniques for the diagnostics of the IMS rolling element bearing dataset. Appl. Sci. Switz. **13**(10) (2023). https://doi.org/10.3390/app13105977

3. Abbasi, M.A., Huang, S., Khan, A.S.: Fault detection and classification of motor bearings under multiple operating conditions. ISA Trans. **156**, 61–69 (2025). https://doi.org/10.1016/j.isatra.2024.11.008
4. Zhou, Q., Yi, C., Yan, L., Huang, C., Lin, J.: A blind deconvolution approach based on spectral harmonics-to-noise ratio for rotating machinery condition monitoring. IEEE Trans. Autom. Sci. Eng. **20**(2), 1092–1107 (2023). https://doi.org/10.1109/TASE.2022.3179457
5. Shi, Z., Liu, J., Xiao, G.: Analysis of cage slip and impact force in a cylindrical roller bearing with race defects. Tribol. Int. **180**, 108208 (2023). https://doi.org/10.1016/j.triboint.2022.108208
6. Ren, J., Zhang, J., Wang, J., Zhao, X.: Recursive signal denoising method for predictive maintenance of equipment by using deep learning based temporal masking. Comput. Ind. Eng. **188**, 109921 (2024). https://doi.org/10.1016/j.cie.2024.109921
7. Bilendo, F., Badihi, H., Lu, N., Jiang, B.: A data-driven prognostics method for explicit health index assessment and improved remaining useful life prediction of bearings. ISA Trans. (2021). https://doi.org/10.1016/j.isatra.2021.05.007
8. Zhong, J., Wang, D., Li, C.: A nonparametric health index and its statistical threshold for machine condition monitoring. Measurement **167**, 108290 (2021). https://doi.org/10.1016/j.measurement.2020.108290
9. Avendano, D.N., Vandermoortele, N., Soete, C.: A semi-supervised approach with monotonic constraints for improved remaining useful life estimation. Sensors (2022). https://www.mdpi.com/1424-8220/22/4/1590
10. Tang, Z., Wang, M., Ouyang, T., Che, F.: A wind turbine bearing fault diagnosis method based on fused depth features in time–frequency domain. Energy Rep. **8**, 12727–12739 (2022). https://doi.org/10.1016/j.egyr.2022.09.113
11. Al Mamun, A., et al.: Multi-channel sensor fusion for real-time bearing fault diagnosis by frequency-domain multilinear principal component analysis. Int. J. Adv. Manuf. Technol. **124**(3–4), 1321–1334 (2023). https://doi.org/10.1007/s00170-022-10525-4
12. Goyal, D., Choudhary, A., Pabla, B.S., Dhami, S.S.: Support vector machines based non-contact fault diagnosis system for bearings. J. Intell. Manuf. **31**(5), 1275–1289 (2020). https://doi.org/10.1007/s10845-019-01511-x
13. Liu, Y., Wang, Y., Wang, Y., Xue, S., Wang, Z., Gao, Z.: Method for predicting remaining useful life of rolling bearings based on dynamic complexity characteristic entropy and quantum neural networks. Eng. Fail. Anal. **170**, 109315 (2025). https://doi.org/10.1016/j.engfailanal.2025.109315
14. Ayman, A., Onsy, A., Attallah, O., Brooks, H., Morsi, I.: Feature learning for bearing prognostics: a comprehensive review of machine/deep learning methods, challenges, and opportunities. Measurement 116589 (2024). https://doi.org/10.1016/j.measurement.2024.116589
15. Jiang, H., et al.: Bearing performance degradation assessment based on SC-RMI and Student's t-HMM. Materials **14**(20), 6077 (2021). https://doi.org/10.3390/ma14206077
16. Patil, A.P., Mishra, B.K., Harsha, S.P.: Fault diagnosis of rolling element bearing using autonomous harmonic product spectrum method. Proc. Inst. Mech. Eng. Part K J. Multi-Body Dyn. **235**(3), 396–411 (2021). https://doi.org/10.1177/1464419321994986
17. Cheng, H., et al.: A novel bearing fault diagnosis method based on improved product envelope spectrum Gini coefficient ratio optimization gram. Mech. Syst. Signal Process. **239**, 113349 (2025). https://doi.org/10.1016/j.ymssp.2025.113349
18. Schmidt, S., Wilke, D.N., Gryllias, K.C.: Generalised envelope spectrum-based signal-to-noise objectives: formulation, optimisation and application for gear fault detection under time-varying speed conditions. Mech. Syst. Signal Process. **224**, 111974 (2025). https://doi.org/10.1016/j.ymssp.2024.111974

19. Fu, Y., Liu, X., Sarkar, S., Wu, T.: Gaussian mixture model with feature selection: an embedded approach. Comput. Ind. Eng. **152**, 107000 (2021)
20. Wardhana, I., Gouiaa-Mtibaa, A., Vrignat, P., Kratz, F.: Contribution to estimating the level of bearing degradation using a Multi-Branch Hidden Markov Model approach. Comput. Ind. **167**, 104254 (2025). https://doi.org/10.1016/j.compind.2025.104254

Research on the Gimballed Inertial Navigation System Based on Mathematical Description of Platform Motion

Olha Sushchenko(✉)

State University "Kyiv Aviation Institute", Liubomyra Huzara Ave., 1, Kyiv 03058, Ukraine
sushoa@ukr.net

Abstract. This paper deals with such a feature of the mathematical description for a gimballed navigation system as the kinematics of a moving platform with mounted navigation instruments. The measuring part of the studied system includes high-precision rate gyroscopes and pendulum accelerometers. The novelty of the proposed approach lies in the full mathematical description of the platform equipped with navigation instruments in the inertial space. This leads to sufficient improvement in the platform's motion control system, synthesised on the basis of the proposed mathematical description. Systems of coordinates that describe the platform's position in the inertial space are presented. The detailed kinematic equations are derived. The expressions for the platforms' angular rates are developed. The features of the parametric optimization of the platform's motion control system are described. The results of modelling the synthesised motion control system in the form of graphical dependencies and calculation of robustness indices are presented. The obtained results can also be useful for the motion control systems of gyro-stabilized platforms with different observation and information-measuring devices.

Keywords: Modelling · Mathematical Description · Motion Control System · Stabilization · Gimballed Navigation System · Parametric Optimization

1 Introduction and Literature Review

Gimballed navigation systems are intended for determining the orientation of a moving object in the inertial space with high accuracy due to the use of gyroscopic devices capable of keeping the orientation in the inertial space unchanged relative to a given axis. Moreover, these gyroscopic devices are mounted on the stabilized platform. The stabilization is implemented by a motion control system, for example, a servo system. Modern approaches to the development of stabilized systems for determining course and attitude are characterized by increasing requirements for their functionality, while at the same time being accompanied by the need to reduce design times. It is possible to reconcile such conflicting requirements only if the system is effectively developed at the initial stages of its creation. One of the key means of ensuring such efficiency is the use of modern methods of optimal synthesis, which, in turn, require the formation of a complete mathematical description of the systems under study.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 37–50, 2027.
https://doi.org/10.1007/978-3-032-30521-3_3

Earlier, such systems were synthesized on the basis of classical methods of given transfer functions and amplitude-frequency characteristics. Nowadays, modern control methods are used, including LQR synthesis, robust synthesis, adaptive synthesis, and methods based on artificial intelligence.

The studied systems are operated on moving objects in difficult conditions of intensive external disturbances. Given the specifics of the systems under study, one should mention methods of robust structural synthesis. H_∞ synthesis and μ synthesis significantly depend on the correspondence of the mathematical description to the real system. Here we are talking about the structure of the model. In addition, H_∞ synthesis is based on a generalized model that takes into account transfer weight functions. The definition of such functions is a complex task based on heuristic approaches. This is crucial for the formation of the amplitude-frequency characteristics of the synthesized system (loop-shaping results). μ synthesis, in addition to the mathematical description of the system, requires mathematical descriptions of uncertainties that significantly affect the results of the synthesis of the motion control system. In other words, the efficiency of robust methods depends essentially on the quality of the system's mathematical description.

When stabilization laws are well known, the acceptable robust approach is the mixed H_2/H_∞ synthesis.

Note that adaptive control systems are resistant to the influence of external disturbances too. Nevertheless, adaptive systems are expensive, complicated to use, and difficult to manufacture. They are practically not used for the platform's stabilization in gimballed navigation systems.

Methods based on artificial intelligence require complex computing units. They are convenient for application in strapdown stabilization systems, where the sequence of rotations in the inertial space due to disturbances is realized by calculations in a computer.

The current stage of developing moving vehicles intended for operation on water requires a radical improvement of navigation and control means of moving objects. The most common such means are considered navigation and stabilization systems based on gyroscopic devices due to their resistance to external influences and a high degree of autonomy [1]. At the same time, serious attention is paid to gyroscopic devices for determining the course of a moving object, primarily gyrocompasses. This is explained by the fact that for reliable and accurate navigation and solving other problems, continuous and fairly accurate information about the current course is necessary.

The development of gimballed systems continues to be relevant when the main goal is to ensure maximum accuracy in autonomous operation or to reduce the duration and cost of creating new systems by using the achievements of previous projects. This trend is especially evident in the field of high-precision systems.

Gimballed inertial navigation systems and features of their application are described in the publication [1]. Such systems are grounded on the inertial navigation [2] and inertially stabilized platform technology [3].

Features of creating models of systems and devices designed to determine the course of moving objects are highlighted in numerous scientific publications.

In particular, in the publication [4], a simplified model of a navigation system using gyroscopes with dynamic tuning is proposed. Similar models are used to study navigation

contours of course determination and stabilization systems. The models of dynamically tuned gyroscopes are used with some additions for studying the gimballed system in the present paper.

In the paper [5], a mathematical model of a small-sized corrected gyrocompass based on a gyroscope with dynamic tuning is presented, which takes into account the movement of the platform. In contrast to the study given in [5], the present paper considers a system with a developed configuration of inertial measuring devices and three-axis gimbals that ensure the possibility of measuring full information about the location of the moving object in the inertial space.

In the paper [6], models of the gimballed course determination system are presented with a division into modes of precise alignment to the horizon and operation in the course measurement mode. However, a platform with two-axis gimbals is studied in this paper. This approach to considering a gimballed navigation system as a multi-mode one is kept and developed in the present paper, taking into consideration the complication of the gimballed platform with inertial measuring instruments for three-axis construction.

A mathematical description of a system of the corresponding type is presented in the paper [7], but it is limited only to the gyroscopic compass mode, without taking into account the platform motion and without a detailed analysis of the control features. The present paper corrects this approach based on a detailed study of the motion of the gimballed platform.

The main goal of the paper is to study the kinematics of the platform with navigation inertial instruments as a main component of a gimballed inertial navigation system. The results of the study can be used in procedures for the synthesis of the system controlling a gimballed platform motion.

The structure of the paper can be represented in the following way. Section 2 deals with the mathematical description of gimballed inertial navigation system kinematics. Section 3 describes the features of the parametric optimization for the gimballed platform motion control system. The results of modelling and calculating indices of the synthesised motion control system are presented in Sect. 4. Conclusions are given in Sect. 5.

2 Mathematical Description of Gimballed Inertial Navigation System Kinematics

The creation of a complete mathematical description of stabilization systems and the determination of the course of dynamic objects will require solving some key tasks, which are listed below:

- It is necessary to recognize the presence of several operating modes, which are differentiated by the warehouse of knowledge, and therefore, are characterized by different features and mathematical descriptions.
- It is important to ensure that the dynamics and kinematics of the platform on which the navigation information sensors are placed is taken into consideration. For this purpose, it is necessary to create consistent models in order to describe the dynamics of the platform and the necessary kinematic relationships.

- It is necessary to consider the modelling and monitoring of the navigation and stabilization contours of the system.
- The analysis of systems may be carried out in accordance with the presence of electronic devices.
- To solve the most significant problems, it is recommended to combine such approaches.

A complete mathematical model of the gimballed navigation system for determining heading and attitude of a moving object is a set of mathematical models that describe the operation of the system in the main modes, including the mode of forward setting to the horizon plane, modes of accurate setting to the horizon plane, gyroscopic compass mode, and gyroscopic azimuth mode.

The system's model in the mode of forward setting to the horizon plane includes the rough alignment of the platform and the alignment of accelerometers. In the mode of precise adjustment to the horizon plane, the mathematical model describes the process of the platform's levelling, the alignment of the gyroscopic vertical based on a dynamically tuned gyroscope, as well as the alignment of pendulum accelerometers.

The most complex models correspond to the gyroscopic compass and gyroscopic azimuth modes. They include equations that describe the platform's dynamics and kinematics, a mathematical description of the gyroscopic vertical, a mathematical description of the gyroscopic compass (gyroscopic azimuth), and mathematical descriptions of accelerometers. The above-mentioned gyroscopic devices are based on such complex units as dynamically tuned gyroscopes. In this case, the peculiarities of controls in the different modes are considered, which are differentiated by the variety of sensors, the main functions, and the ability to achieve accuracy.

The mathematical description of the platform's motion can be obtained on the basis of the theorem about the change in the kinetic moment of a solid body, which will be described below. In this case, it is necessary to completely restore all kinematic correlations of the system. It should be appreciated that such an approach leads to very complex mathematical expressions. This approach can be simplified using a simpler method, which does not account for the difference between the moments of inertia of the different axes of the gimbals.

To extract kinematic relations, it is necessary to introduce the following coordinate systems.

The coordinate system $Ox_0y_0z_0$ is associated with the moving object.

The coordinate system $Ox_py_pz_p$ is linked to the platform. It is called the body-axis coordinate system.

The transition from the coordinate system associated with the moving object to the coordinate system associated with the platform is indicated by the subsequent turns. These rotations are implemented around the axis of the outer frame of the platform's gimbal (angular position α), around the inner frame of the platform's gimbal (angular position β), and the platform's rotation (angular position γ). The designated sequence of turns is shown in Fig. 1.

Note that in such a coordinate system at the initial position, the outer axis of the gimbals rotates together with the longitudinal axis of the moving object, and the inner axis of the gimbals with the lateral axis of the moving object.

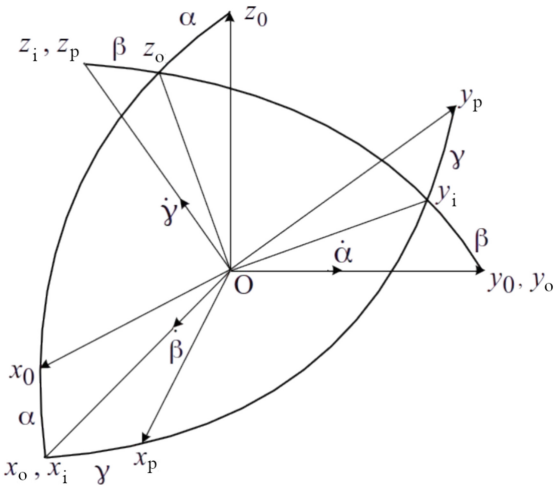


Fig. 1. The relative position of the coordinate systems associated with the moving object and the platform.

The equations of motion of the platform can be derived based on the theorem on the change of angular momentum of a rigid body, which is written below [8–10].

$$\begin{aligned}
&\frac{dK_{zp}}{dt} + \omega_{xp}K_{yp} - \omega_{yp}K_{xp} = M_{zp};\\
&\frac{dK_{xi}}{dt} + \omega_{yi}K_{zi} - \omega_{zi}K_{yi} = M_{xi};\\
&\frac{dK_{yo}}{dt} + \omega_{zo}K_{xo} - \omega_{xo}K_{zo} = M_{yo},
\end{aligned} \tag{1}$$

where K_{xp}, K_{yp}, K_{zp}, ω_{xp}, ω_{yp}, ω_{zp} are the kinetic moments and angular velocities of the platform with the installed gyroscopic devices and pendulum accelerometers; M_{xp}, M_{yp}, M_{zp} are external moments acting on the rotating platform. These moments are divided into controls and disturbances. Controls are defined by the mode of the gimballed navigation system. Disturbances depend on the type of moving object.

The kinetic moments included in the theorem on the change of angular momentum of a rigid body (1) are calculated using the following formulas describing the kinematics of the platform.

$$\begin{aligned}
K_{xp} &= I_{xp}\omega_{xp};\\
K_{yp} &= I_{yp}\omega_{yp};\\
K_{zp} &= I_{zp}\omega_{zp},
\end{aligned} \tag{2}$$

where I_{xp}, I_{yp}, I_{zp} are moments of the platform with devices.

The kinematics of the inner gimbal frame can be represented by relationships.

$$\begin{aligned}
K_{xi} &= I_{xi}\omega_{xi} + K_{xp}\cos\gamma - K_{yp}\sin\gamma;\\
K_{yi} &= I_{yi}\omega_{yi} + K_{yp}\cos\gamma + K_{xp}\sin\gamma;\\
K_{zi} &= I_{zi}\omega_{zi} + K_{zp},
\end{aligned} \tag{3}$$

where $K_{xi},\ K_{yi},\ K_{zi}, I_{xi},\ I_{yi},\ I_{zi}, \omega_{xi},\ \omega_{yi},\ \omega_{zi}$ are the corresponding kinetic moments, moments of inertia, and angular velocities of the inner frame of the gimballed platform; γ is the angle of the platform's rotation.

The mathematical description of the kinematics of the outer gimbal frame are.

$$\begin{aligned} K_{xo} &= I_{xo}\omega_{xo} + K_{xo}; \\ K_{yo} &= I_{yo}\omega_{yo} + K_{yo}\cos\beta - K_{zo}\sin\beta; \\ K_{zo} &= I_{zo}\omega_{zo} + K_{zo}\cos\beta + K_{yo}\sin\beta, \end{aligned} \tag{4}$$

where $K_{xo},\ K_{yo},\ K_{zo}, I_{xo},\ I_{yo},\ I_{zo}, \omega_{xo},\ \omega_{yo},\ \omega_{zo}$ are the kinetic moments, moments of inertia, and angular velocities of the outer frame of the gimballed platform; β is the angle of the inner frame rotation.

By substituting the expressions for kinetic moments (2) – (4) into the equation of motion of the platform (1), we obtain an equation describing the motion of the platform relative to its own axes

$$I_{zp}\dot{\omega}_{zp} + I_{yp}\omega_{xp}\omega_{yp} - I_{xp}\omega_{xp}\omega_{yp} = M_{zp}. \tag{5}$$

After the conversion of Eq. (5), we will obtain

$$I_{zp}\dot{\omega}_{zp} + \left(I_{yp} - I_{xp}\right)\omega_{xp}\omega_{yp} = M_{zp}. \tag{6}$$

The equations of motion of the platform relative to the inner axis of the gimbals become

$$\begin{gathered} I_{xi}\dot{\omega}_{xi} + I_{xp}\dot{\omega}_{xp}cos\gamma - I_{yp}\dot{\omega}_{yp}sin\gamma + I_{zi}\omega_{yi}\omega_{zi} + I_{zp}\omega_{yi}\omega_{zp} - \\ -I_{yi}\omega_{yi}\omega_{zi} - I_{yp}\omega_{zi}\omega_{yp}cos\gamma - I_{xp}\omega_{zi}\omega_{xp}sin\gamma = M_{xi}. \end{gathered} \tag{7}$$

After conversion of the Eq. (7), we will obtain:

$$\begin{gathered} I_{xi}\dot{\omega}_{xi} + \left(I_{zi} - I_{yi}\right)\omega_{yi}\omega_{zi} + I_{zp}\omega_{yi}\omega_{zp} + \\ +\left(I_{xp}\dot{\omega}_{xp} - I_{yp}\omega_{zi}\omega_{yp}\right)cos\gamma - \left(I_{yp}\dot{\omega}_{yp} + I_{xp}\omega_{zi}\omega_{xi}\right)sin\gamma = M_{xi}. \end{gathered} \tag{8}$$

The equations of motion of the platform relative to the external axis of the gimbal are described as follows:

$$\begin{aligned} &I_{yo}\dot{\omega}_{yo} + I_{yi}\dot{\omega}_{yi}\cos\beta + I_{yi}\dot{\omega}_{yp}\cos\gamma\cos\beta + I_{xp}\dot{\omega}_{xp}\sin\gamma\cos\beta - I_{zi}\dot{\omega}_{zi}\sin\beta - \\ &-I_{zp}\dot{\omega}_{zp}\sin\beta + (I_{xo} - I_{zo})\omega_{xo}\omega_{zo} + I_{xi}\omega_{xi}\omega_{zo} + I_{xp}\omega_{xp}\omega_{zo}\cos\gamma - \\ &-I_{yp}\omega_{zo}\omega_{yp}\sin\gamma - I_{zi}\omega_{xo}\omega_{zi}\cos\beta - I_{zp}\omega_{xo}\omega_{zp}\cos\beta - \\ &-I_{yi}\omega_{xo}\omega_{yi}\sin\beta - I_{yp}\omega_{yp}\omega_{xo}\cos\gamma\sin\beta - I_{xp}\omega_{xo}\omega_{xp}\sin\gamma\sin\beta = M_{yo} \end{aligned} \tag{9}$$

As shown in Fig. 1, the expressions for the angular velocities of the platform can be presented in the following form:

for the outer gimbals frame

$$\begin{aligned} \omega_{xo} &= \omega_{x0}\cos\alpha - \omega_{z0}\sin\alpha, \\ \omega_{yo} &= \dot{\alpha} + \omega_{y0}, \\ \omega_{zo} &= \omega_{z0}\cos\alpha + \omega_{x0}\sin\alpha, \end{aligned} \tag{10}$$

for the inner gimbals frame

$$\begin{aligned}\omega_i &= \dot{\beta} + \omega_{xo},\\ \omega_{yi} &= \omega_{yo}\cos\beta + \omega_{zo}\sin\beta,\\ \omega_{zi} &= \omega_{zo}\cos\beta - \omega_{yo}\sin\beta,\end{aligned} \tag{11}$$

for the platform

$$\begin{aligned}\omega_{xp} &= \omega_{xi}\cos\gamma + \omega_{yi}\sin\gamma,\\ \omega_{yp} &= \omega_{yi}\cos\gamma - \omega_{xi}\sin\gamma,\\ \omega_{zp} &= \dot{\gamma} + \omega_{zi},\end{aligned} \tag{12}$$

where $\omega_{x0},\ \omega_{y0},\ \omega_{z0}$ are components of the object's velocity; $\omega_{xp},\ \omega_{yp},\ \omega_{zp}$ are components of the platform's velocity; $\omega_{xi},\ \omega_{yi},\ \omega_{zi}$ are projections of the platform on the axis of the coordinate system associated with the internal axis of the gimbal; $\omega_{xo},\ \omega_{yo},\ \omega_{zo}$ are projections of the platform on the axis of the coordinate system associated with the external axis of the gimbals.

After the substitution of relationships (6), (8), (9) in expressions (10) – (12), the equations of motion of the platform can be presented in a complete form. Due to the complexity and cumbersomeness of the transformations, only the equation of motion of the platform relative to its own axis is given in the work. To achieve the maximum possible simplification, a special case of the motion of an object with a constant course, in the small-angle regime, was considered. The resulting equation has the following form:

$$\begin{aligned}&\ddot{\gamma} + \tfrac{\dot{\alpha}}{I_{zp}}\omega_{x0} + \tfrac{I_{yp}-I_{xp}}{I_{z\Pi}}[cos2\gamma\left(\dot{\alpha}\omega_{x0} + \dot{\beta}\omega_{z0} + \dot{\beta}\omega_{y0} + \omega_{z0}\omega_{x0} + \omega_{y0}\omega_{x0}\right)+\\ &+0.5sin2\gamma\left(2\dot{\alpha}\omega_{y0} + \omega_{y0}^2 - 2\dot{\beta}\omega_{x0} - \omega_{x0}^2 + 2\omega_{x0}\omega_{z0}\alpha + 2\omega_{y0}\omega_{z0}\beta\right)] = M_{zp}.\end{aligned} \tag{13}$$

The projections of the angular velocity of an object can be determined using the following relationship:

$$\omega_0 = \omega_{0'} + \overset{''}{\omega_0}, \tag{14}$$

where $\omega_{0'}$ is the velocity of the initial reference frame $O\xi\eta\zeta$; ω_0'' is the velocity of the body relative to this reference frame.

The position of the body relative to the initial reference frame, which is determined by rotations through angles $\psi_o,\ \vartheta_o,\ \varphi_o$, is shown in Fig. 2. In accordance with the representation given in Fig. 2, the formulas for velocity projections $\omega_{0'}$ as a component in the expression (14) can be presented in the following form:

$$\begin{aligned}\omega_{x0'} &= -\dot{\psi}_o\cos\vartheta_o\sin\varphi_o + \dot{\vartheta}_o\cos\varphi_o\\ \omega_{y0'} &= \dot{\varphi}_o + \dot{\psi}_o\sin\vartheta_o,\\ \omega_{z0'} &= \dot{\psi}_o\cos\vartheta_o\cos\varphi_o + \dot{\vartheta}_o\sin\varphi_o\end{aligned} \tag{15}$$

To complete the full mathematical description, it is necessary to represent the relationships for the description of the angular rate using parameters defined in the system of coordinates associated with the moving object.

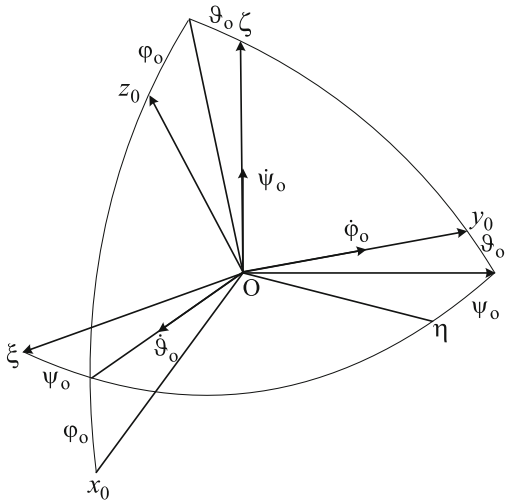


Fig. 2. Location of the object relative to the initial navigation coordinate system.

The motion of an object relative to the initial coordinate system ω_0'' as a component of the expression (14) can be described using the following expressions:

$$\begin{aligned}
\omega_{x0}'' &= -\frac{V_{xp}}{R_1}\sin k - \frac{V_{yp}}{R_2}\cos k - \Omega\cos\varphi\sin k;\\
\omega_{y0}'' &= \frac{V_{xp}}{R_1}\cos k + \frac{V_{yp}}{R_2}\sin k + \Omega\cos\varphi\cos k;\\
\omega_{z0}'' &= \frac{V_{xp}}{R_1}\mathrm{tg}\phi - \dot{k} + \Omega\sin\varphi,
\end{aligned} \tag{16}$$

where: Ω is the speed of the Earth's motion; φ is the geographical latitude of the point; $R_1 = \frac{a\cos\varphi}{\sqrt{1-e^2\sin^2\varphi}}$ is the principal radius of curvature of the Earth's ellipsoid in the plane perpendicular to the meridian plane; $R_2 = \frac{a(1-e^2)}{(1-e^2\sin^2\varphi)^{3/2}}$ is the principal radius of curvature of the Earth's ellipsoid in the meridional plane; a, b are the semi-major and minor axes of the Earth's ellipsoid; $e = \frac{\sqrt{a^2-b^2}}{a}$ is the magnitude of the eccentricity; k is the direction of motion of the object.

Other features of the full mathematical description of the studied system, including mathematical models of gyroscopic devices, accelerometers, and electronic units, are presented in the publication [11].

3 Features of Parametric Optimization

To develop a computational algorithm for parametric optimization of the system, it is necessary to take into account its specific characteristics. In the case of the stabilized attitude and heading system, these features are as follows. First of all, it is a static second-order system, which necessitates the need to perform its minimum implementation before

optimizing the parameters. A study of the system matrix of the model of this system shows that the values of its elements differ by approximately three orders of magnitude. Given this, it is advisable to use a balanced implementation of the model. As parameters to be optimized, it is advisable to choose the transmission coefficients that determine the control laws. The parametric optimization process is implemented in two consecutive stages. The first stage involves the direct implementation of parameter optimization using the appropriate models developed using the advanced tools of the Control System Toolbox and Simulink Toolbox. The second stage involves the verification of the obtained results using a mathematical model that reflects all the nonlinear properties inherent in real systems and, if possible, includes models of all components that are part of the system under study.

The essence of the parametric optimization of the studied system lies in ensuring resistance to the internal and external disturbances. Such an approach can be based on the so-called mixed approach when the optimization criterion using H_2, H_∞ norms is used [12–15]. The important component of the optimization criterion is the penalty function, which checks the stability of the motion control system during the optimization procedure. The algorithm of the parametric disturbances in application to the studied systems, taking into consideration the necessity of the precise stabilization of the platform with measuring instruments, is represented in publications [16, 17]. The synthesis of systems of another class, which are resistant to disturbances are given in publication [18–21]. These paper describe features of the robust control for ground and aerial vehicles. Some general aspects proposed in these publications were useful for developing the algorithm of the robust optimization for studied in the present paper system assigned for operation on marine vehicles with the necessity of realizing precise navigation measurements.

4 Modelling Results

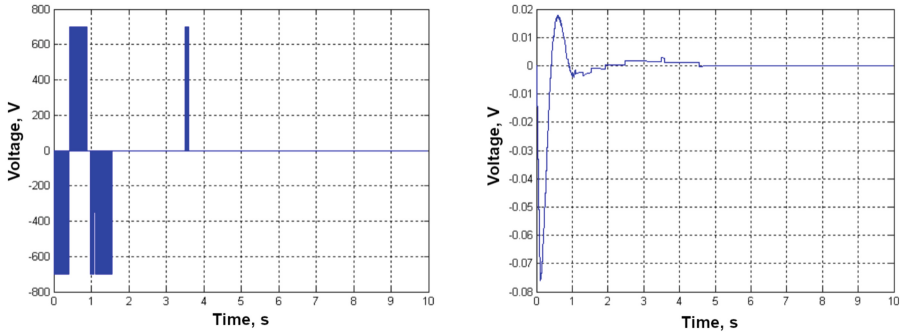
The modelling of the studied system has been carried out in MATLAB environment and software, which is well adapted to solving problems associated with creating complex control systems, including those resistant to disturbances based on robust and adaptive approaches [22, 25]. These books are well known among scientists, which deal with the optimal synthesis of motion control systems of the wide class. They include many useful examples that show possibilities of MATLAB in realization of automatic control systems of the wide class.

The application of MATLAB to a successful solution of structural robust control for stabilization observation equipment on unmanned aerial vehicles is represented in papers [26, 27]. The present paper deals with parametric robust optimization. Nevertheless, ideas represented in [26, 27] were taken into consideration for creating the algorithm of platform stabilization in conditions of uncertainty. Examples of modelling for ground-moving vehicles, drive control systems, and aerial stochastic dynamic objects with unspecified dynamics are carried out in [28, 29]. Some similarity in the control of these objects and the studied moving platforms allowed us to improve some aspects of modelling.

For the studied system, simulation in different modes of the gimballed inertial navigation system has been carried out. The results of modelling optimal control laws in

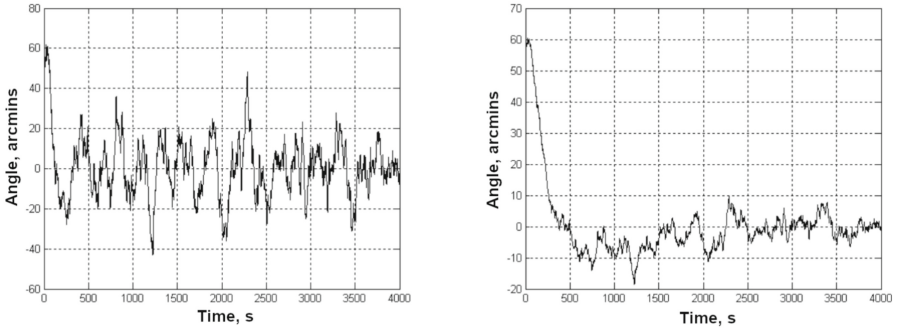
the mode of preliminary horizon reduction, performed using a mathematical model approximated to the real gimballed system motion, are shown in Figs. 3 and 4. The obtained optimal control laws are based on a stochastic computational procedure taking into consideration the complex motion of the platform with the kinematics described by Eqs. (13), (15), (16).

A comparison of the stability of optimized and non-optimized systems to the action of disturbances in the stabilization mode is shown in Fig. 4.



a b

Fig. 3. Output signals of the pulse-width modulator (a) and calculating unit (b).



a b

Fig. 4. Transient processes of changing angles of deviation from the horizontal plane in a perturbed non-optimized system (a) and optimised (b) systems.

It is obvious that the optimized system in real operating conditions shows better stabilization accuracy (see Fig. 4b). A characteristic feature of parametric optimization of the stabilized attitude and course determination system is the use of H_2, H_∞ norms for both channels with the weighting coefficients, which are taken into account on equal terms.

The corresponding norms of the synthesized system are 0.1224 and 0.0054, respectively. These values (sufficiently less than 1) prove the low root mean square error and high robustness (stability to disturbances) of the synthesised system, correspondingly.

Consider the results of modelling in the mode of setting to the meridian. The process of parametric optimization of the studied system is accompanied by obtaining a minimal and balanced implementation of the model. In this mode, the H_2, H_∞ norms, which characterize the accuracy and robustness of the system, were improved sufficiently from 7.24 to 0.78 and from 2.81 to 0.08, respectively. These values prove the high efficiency of the parameter optimization from the point of view of high accuracy achievement (defined by the low value of H_2 norm) and disturbance suppression (defined by the low value of H_∞ norm). The high resistance to the disturbance for the synthesised system is explained also by the additional means directed to achieving the same goal, for example, application of the integral correction ensuring realization of the undisturbed gyroscopic vertical.

It should be noted that, in accordance with the purpose of the studied type of system, the most significant component of the control law is the one responsible for the accelerated drive to the meridian. It is advisable to include restrictions on this coefficient in the penalty function, since this allows you to control the speed of the process of bringing to the meridian. At the same time, reducing the time of bringing to the meridian is accompanied by an increase in the oscillations and norms of the system. The results of parametric optimization for the limiting values of this coefficient are shown in Fig. 5. It shows the transient process in the azimuth angle for the minimum (a) and maximum (b) values of the appropriate coefficient. In both cases, the stability margin in amplitude is 49.6 dB, and in phase, 76°. At the same time, there is a significant difference in the time of alignment to the meridian: 943 s and 601 s, respectively.

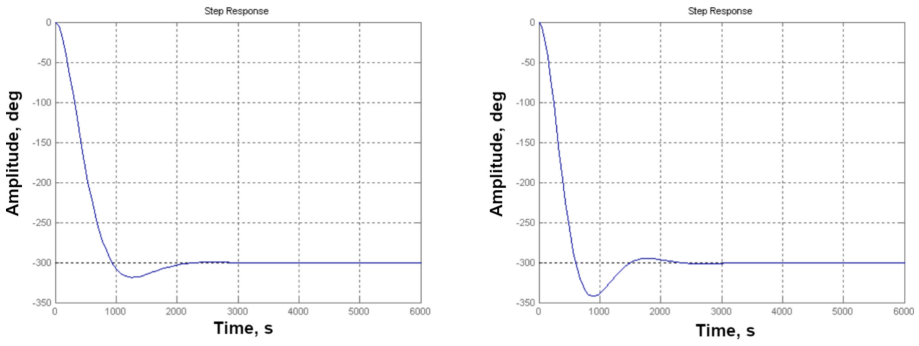


Fig. 5. Transients in the mode of setting to the meridian in the usual and accelerated modes.

The features of using MATLAB computing platform for the synthesis and modelling of the motion control systems studied in this paper are given in the books [22–24]. During modelling, Control System Toolbox and Robust Control Toolbox have been used. The full descriptions of these Toolboxes are given in the User's Guides [25, 30].

5 Discussion and Conclusions

The research aims to improve the operating characteristics of the gimballed inertial navigation systems using the detailed mathematical description of the stabilized platform kinematics for the synthesis of the high-precision gimballed navigation system.

The main achievement is taking into consideration the gimballed platforms' motion in the mathematical description of the gimballed navigation systems. This leads to sufficient improvement of the platform's stabilization and, hence, navigation accuracy. The obtained results can be useful for the creation of robust control systems in many areas of application of moving objects, including aerial, marine, and ground ones. Nevertheless, the main contribution of the obtained results can be achieved for the development of precision navigation systems.

Finally, it is possible to represent the following conclusions.

In the presented paper, the mathematical description of the kinematics for the gimballed inertia navigation system is presented. The coordinate systems for the complex motion of the platform with navigation instruments, which ensure its motion by three degrees of freedom, are represented. The equation of the platform's motion is represented. Expressions of the platform's angular rates are derived. The developed mathematical description of the complex motion of the gimballed platform with navigation instruments such as gyroscopic devices and accelerometers allows us to take into consideration the platform's complex motion of the platform and increase the accuracy of the motion control system synthesised based on the proposed mathematical description.

The features of the parametric optimization of motion control systems resistant to internal and external disturbances are generalized and described. The appropriate optimization criterion is analysed.

The modelling results, taking into consideration different aspects of the multi-mode gimballed inertial navigation system, such as previous levelling, stabilization, and setting to the meridian, are given. The representation of transient processes and calculation of H_2, H_∞ norms is given.

The presented results can be useful for applications that require precision stabilization based on the possibility of controlling a moving platform with highly accurate navigation instruments. At the same time, the obtained results can be applicable for the precision stabilization of various measurement and observation devices (antennas, cameras, and so on), which nowadays are widely used on moving objects of different types.

The future work will be connected with a similar mathematical description of the gimballed system assigned for the stabilization of the observation equipment mounted on moving objects operated in difficult, disturbed conditions. The emphasis will be given on the instrumentation of the ground-moving objects.

Acknowledgments. This study is partially supported by the Ministry of Education and Science of Ukraine in the framework of the scientific research work "Methodology of High-Precision Stabilized Platforms for Aircraft Equipment using Perspective Inertial Technologies" (№ 0119U100654).

References

1. Wang, H.G., Williams, T.C.: Strategical inertial navigation systems. IEEE Control. Syst. Mag. **1**, 65–85 (2008). https://doi.org/10.1109/MCS.2007.910206
2. Jekeli, C.: Inertial Navigation Systems with Geodetic Applications. De Gruyter, Berlin (2023). https://doi.org/10.1515/9783110784329
3. Hilkert, J.M.: Inertially stabilized platform technology. IEEE Control. Syst. Mag. **1**, 26–46 (2008). https://doi.org/10.1109/MCS.2007.910256
4. Xia, D., Yu, C., Kong, L.: A micro dynamically tuned gyroscope with adjustable static capacitance. Sensors **13**(2), 2176–2195 (2013). https://doi.org/10.3390/s130202176
5. Zamorsky, A.: Gyroscope with indirect control for ground vehicles. Bulletin of Kyiv Polytechnic Institute, Ser. Instrument Making **59**(1), 5–23 (2020). https://doi.org/10.20535/1970.59(1).2020.210165
6. Sushchenko, O., et al.: Algorithm of determining errors of gimballed inertial navigation system. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., C. Rocha, A.M.A., Faginas Lago, M.N. (eds) Computational Science and Its Applications – ICCSA 2024 Workshops, Lecture Notes in Computer Science, vol. 14816, pp. 206–218. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-65223-3_14
7. Jia, R., Nandikolla, V., Haggart, G., Volk, C., Tazartes, D.: System performance of an inertially stabilized gimbal platform with friction, resonance, and vibration effects. J. Nonlinear Dyn. ID **6594861**, 1–20 (2017). https://doi.org/10.1155/2017/6594861
8. Mukhina, M.P., Filyashkin, M.K.: Computer modeling of accuracy characteristics of strapdown inertial navigation system. Radio Electron. Comput. Sci. Control (4), 157–165 (2019). https://doi.org/10.15588/1607-3274-2019-4-15
9. Braasch, M.: Fundamentals of Inertial Navigation Systems and Aiding. IET, London (2022). https://doi.org/10.1049/SBRA550E
10. Niu, S.S., Xiao, D.: Process Control: Engineering Analysis and Best Practices. Springer, New York (2022)
11. Sushchenko, O., Bezkorovainyi, Y., Salyuk, O., Yehorov, S.: Mathematical description of system for stabilization of aviation equipment in problems of synthesis and simulation. In: Ostroumov, I., Zaliskyi, M. (eds) Proceedings of the 2nd International Workshop on Advances in Civil Aviation Systems Development. ACASD 2024, Lecture Notes in Networks and Systems, vol. 992, pp. 73–85. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-60196-5_7
12. Scogestad, S., Postlathwaite, I.: Multivariable Feedback Control. Analysis and Design. John Wiley & Sons, London (2005)
13. Fu, J., Ma, R.: Stabilization and Hinf Control of Switched Dynamic Systems. Springer, Berlin (2020)
14. Zerrik, El. H., Castillo O.: Stabilization of Infinite Dimensional Systems. Springer, Cham (2021)
15. Sanfelice, R.G.: Hybrid Feedback Control. Princeton University Press, Princeton (2021)
16. Sushchenko, O.A., Shyrokyi, O.V.: H2/H∞ optimization of system for stabilization and control by line-of-sight orientation of devices operated at UAV. In: IEEE 3rd International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), pp. 235–238. IEEE, Kyiv, Ukraine (2015). https://doi.org/10.1109/APUAVD.2015.7346608
17. Sushchenko, O.A.: Robust control of angular motion of platform with payload based on H∞-synthesis. J. Autom. Inf. Sci. **48**(12), 13–26 (2016). https://doi.org/10.1615/JAutomatInfScien.v48.i12.20
18. Kuznetsov, B.I., Nikitina, T.B., Bovdui, I.V.: Multiobjective synthesis of two degree of freedom nonlinear robust control by discrete continuous plant. Tech. Electrodynamics **5**, 10–14 (2020). https://doi.org/10.15407/techned2020.05.010

19. Voliansky, R. et al.: Interval modeling and simulation of duffing pendulum. In: IEEE 4th KhPI Week on Advanced Technology (KhPIWeek), pp. 1–6. IEEE, Kharkiv, Ukraine (2023) https://doi.org/10.1109/KhPIWeek61412.2023.10312997
20. Le, A.T.: Adaptive Robust Control Systems. Intech Open, Vienna (2018)
21. Sushchenko, O.: Design of robust tracking system for combined control of information and measuring devices of aircraft. In: Ostroumov, I., Marais, K., Zaliskyi, M. (eds) Advances in Civil Aviation Systems Development. ACASD 2025, Lecture Notes in Networks and Systems, vol. 1418, pp. 389–400. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-91992-3_26
22. Wang, L.: PID Control System Design and Automatic Tuning using MATLAB/Simulink. Wiley, Hoboken (2020)
23. Asadi, F.: State-Space Control Systems: The MATLAB/Simulink Approach. Morgan & Claypool, Sun Rafael (2021)
24. .Gu, D.W., et al.: Robust Control Design with MATLAB. Springer-Verlag, London (2005)
25. Balas, G., Chiang, R., Packard, A., Safonov, M.: Robust Control Toolbox User's Guide. The Math Works Inc., Natick, Massachusetts (2008)
26. Sushchenko, O., Goncharenko, A.: Design of robust systems for stabilization of unmanned aerial vehicle equipment. Int. J. Aerospace Eng. ID **6054081**, 1 (2016). https://doi.org/10.1155/2016/6054081
27. Voliansky, R., Sadovoi, A., Volianska, N.: Interval model of the piezoelectric drive. In: 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), pp. 1–6. IEEE, Lviv-Slavske, Ukraine (2018). https://doi.org/10.1109/TCSET.2018.8336211
28. Osadchy, S.I., Zozulya, V.A., Bereziuk, I.A., Melnichenko, M.M.: Stabilization of the angular position of hexapod platform on board of a ship in the conditions of motions. Autom. Control. Comput. Sci. **56**(3), 221–229 (2022). https://doi.org/10.3103/S0146411622030051
29. Osadhyi, S., Zozulia, V.: Synthesis of optimal multivariable robust systems of stochastic stabilization of moving objects. In: 5th International Conference Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), pp. 106–111. IEEE, Kyiv, Ukraine (2019). https://doi.org/10.1109/APUAVD47061.2019.8943861
30. Control System Toolbox User's Guide. The Math Works Inc., Natick, Massachusetts (2024). https://www.mathworks.com/help/control/

Adaptive Sampling Design for Bayesian Spatial Risk Mapping of Maritime Surveillance in Central Portugal

Belchior Miguel[1,2(✉)], Isabel Natário[1,3], Rui Gonçalves de Deus[4,5], and Paula Simões[1,6,7,8]

[1] Center for Mathematics and Applications (NOVA Math),NOVA University of Lisbon, Lisbon, Portugal
b.miguel@alumni.fct.unl.pt, {icn,pc.simoes}@fct.unl.pt
[2] Instituto Superior Politécnico de Tete, Tete, Mozambique
[3] Department of Mathematics of NOVA School of Science and Technology, Caparica, Portugal
[4] Centro de Investigação Naval (CINAV), Instituto Universitário Militar, Lisbon, Portugal
[5] Direção de Análise e Gestão da Informação, Marinha Portuguesa, Lisbon, Portugal
goncalves.deus@marinha.pt
[6] Centro de Investigação, Desenvolvimento e Inovação da Academia Militar (CINAMIL), Lisbon, Portugal
[7] Instituto Universitário Militar, Lisbon, Portugal
[8] Department of Mathematics, Lisbon Institute of Engineering (ISEL), Polytechnic University of Lisbon, Lisbon, Portugal

Abstract. The Portuguese Navy faces the important operational challenge of contributing to more effective and coordinated management of inspection and control across the Exclusive Economic Zone and national borders, as part of its mission to ensure maritime surveillance and enforce the law.

From an operational perspective, patrol activity tends to concentrate in areas characterised by intense maritime traffic, proximity to ports, and zones traditionally associated with fishing activities. The spatial distribution of inspected events or detected infringements reflect not only the pressure exerted by maritime activity but also the intensity and frequency of enforcement actions. Consequently, the spatial distribution of enforcement operations and recorded infringements contributes both the level of maritime activity and the surveillance strategy adopted. Within this framework, factors such as distance from the coastline, proximity to ports, and the spatial extent of the Exclusive Economic Zone (EEZ) act as structural drivers of the variability observed in enforcement data. Using geostatistical methods combined with INLA (Integrated Nested Laplace Approximation) approach jointly with SPDE (Stochastic Partial Differential Equations) formulation, this study explores the spatial distribution of suspected fishing infractions along the central coast of Portugal.

Some sampling design approaches are proposed for the surveillance of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 51–68, 2027.
https://doi.org/10.1007/978-3-032-30521-3_4

fishing activities, providing adequate support for the Portuguese Navy's maritime surveillance work, thereby supporting efficient and properly coordinated protection of the maritime domain and borders.

Keywords: Bayesian geostatistics · Presence/absence spatial data · Adaptive sampling design · Fisheries-related infractions · Maritime surveillance · Risk mapping

1 Introduction

The surveillance and control of maritime areas under Portuguese sovereignty and jurisdiction are based on an integrated institutional framework combining naval capabilities, maritime authority, and specialised policing. This structure aims to ensure the State's presence at sea, guarantee navigational safety, protect marine resources, and enforce legislation applicable to maritime activities, including fisheries, transport, recreational boating, and resource exploitation.

The naval component is ensured by the Portuguese Navy, which operates assets specifically designed for patrol, inspection, and support to public-interest operations. Among these, the offshore patrol vessels stand out, as they are designed to operate for extended periods at sea and to cover large areas of the Exclusive Economic Zone (EEZ). These vessels perform State authority functions at sea, including surveillance, inspection, support to search and rescue (SAR) operations, and cooperation with other national and international entities [1,2].

In parallel, within the framework of the Maritime Authority System, the National Maritime Authority and the Maritime Police exercise inspection and specialised policing powers in maritime areas, port zones, and the maritime public domain. Their actions focus on ensuring compliance with legal and regulatory provisions, contributing to maritime safety, environmental protection, and the orderly management of activities conducted along the coastline and within the territorial sea [2].

At the coastal level, the Coastal Control Unit of the National Republican Guard complements this framework through surveillance, patrol, and interception activities, both on land and within the territorial sea. This institutional coordination enables broader operational coverage and strengthens the capacity to respond to infringements, irregularities, and risk situations [3].

From an operational perspective, patrol activity tends to concentrate in areas characterised by intense maritime traffic, proximity to ports, and zones traditionally associated with fishing activities. Consequently, the spatial distribution of inspected events or detected infringements may reflect not only the pressure exerted by maritime activity but also the intensity and frequency of enforcement actions. This aspect is particularly relevant in quantitative and spatio-temporal modelling, where event occurrence may be associated both with exposure to risk and with the probability of inspection.

The organisational and operational structure of the Portuguese maritime system thus reflects an integrated model of ocean governance, combining strategic naval presence, specialised policing, and interinstitutional coordination, with

objective of ensuring State authority, maritime safety, and the sustainability of economic activities conducted within national maritime space [4,5].

This study explores the spatial distribution of suspected fishing-related infractions along the central coast of Portugal using geostatistical methods. Based on georeferenced binary observations (presence/absence), the analysis seeks to estimate and visualise the probability of infraction occurrence across the study area. In addition, the study evaluates sampling design strategies derived from geostatistical modelling specifically the Maximum Risk Sampling Design (MaxRSD) and the Maximum Variance Risk Sampling Design (MaxVRSD) to support the analysis of binary data on fishing-related infractions along the Portuguese coast.

Regression models for binary response variables are widely used in many scientific applications, with logistic regression representing one of the most common modelling approaches [6]. These models fall within the class of generalized linear models (GLMs), which relate the expected value of the response variable to a set of covariates through an appropriate link function [7].

In a spatial framework, binary geostatistical data represent observations indicating the presence or absence of a phenomenon at specific locations, allowing the study region to be partitioned into areas where the event occurs and areas where it does not [8]. Spatial dependence in such data is commonly addressed using hierarchical Bayesian spatial models, where spatial effects are introduced through latent Gaussian random fields incorporated into the linear predictor, leading to generalized linear mixed models.

In this study, geostatistical modelling for presence/absence data is implemented using the INLA (Integrated Nested Laplace Approximation) approach combined with SPDE (Stochastic Partial Differential Equations) formulation. This framework provides an efficient representation of the spatial random field and enables computationally efficient Bayesian inference. The models are fitted using the R-INLA framework, allowing the estimation of the spatial field (posterior mean and standard deviation) and the construction of spatial risk maps associated with the analysed infractions [9].

A related study was previously conducted by Miguel et al. [10] for the Algarve region. The present work extends this research by using new datasets, additional explanatory variables, and updated inspection routes, allowing the proposed models to be evaluated in a different geographical context. Furthermore, new sampling design criteria are introduced and compared with a stratified sampling design. The predictive performance of the competing approaches is assessed using RMSE, Log-Score, and LGOCV.

The remainder of the paper is structured as follows. Section 2 outlines the theoretical background of geostatistical models for binary data. Section 3 presents the modelling framework applied to the maritime surveillance data, including the spatial models and the proposed adaptive sampling design criteria, together with the corresponding results. Section 4 discusses the main findings and concludes the study.

2 Sampling Designs for Surveillance Fishing Activities

Sampling design plays a fundamental role in spatial studies because the configuration of sampling locations directly influences both parameter estimation and spatial prediction. Two main paradigms are commonly considered in spatial sampling: the *design-based* approach and the *model-based* approach [11,12].

In the design-based approach, inference relies on the randomization embedded in the sampling design. The population values are considered fixed but unknown, and the validity of statistical inference depends on the probabilistic properties of the sampling scheme. This approach is typically preferred when the objective is to estimate global or regional characteristics of a population, such as spatial cumulative distribution functions or averages over predefined areas, where statistical validity is prioritized over efficiency.

In contrast, the model-based approach treats the spatial phenomenon as a realization of an underlying stochastic process [12]. Within geostatistics, it assumes that the observed spatial data correspond to a single realization of an underlying spatial process defined over the study region [13]. The only source of stochasticity is the assumed spatial process itself, and inference depends on the adequacy of the specified model. This approach is particularly suitable when the objective is to produce spatial maps or predict values at unsampled locations, where predictive accuracy and efficiency are of primary importance.

Sampling design within this framework often involves selecting sampling locations in order to optimize a given objective function [14,15]. Common criteria include the minimization of prediction variance, the minimization of estimation variance for model parameters, or the improvement of spatial prediction accuracy [16]. Alternative strategies have also been proposed that do not require prior knowledge of the covariance structure, focusing instead on improving the estimation of the semivariogram, which is a key component of geostatistical analysis [14,17,18]. In practice, preliminary data exploration or empirical semivariogram estimates may guide the design without requiring a fully specified covariance model.

Although many optimal design approaches assume that the parameters governing spatial dependence are known, this assumption is often unrealistic in practice. When little or no prior information about the spatial structure is available, preliminary data may be used to estimate the covariance structure before refining the sampling design. This flexibility is particularly important in applications where the spatial dependence of the phenomenon is initially unknown.

Importantly, probability-based sampling designs can also be applied when the variable of interest is binary, such as in prevalence studies or in the detection of rare events [19]. In such cases, the spatial distribution of occurrences (e.g., presence or absence of an event) can be analyzed within a geostatistical modelling framework.

Under the model-based paradigm adopted in this work, the sampling locations define the spatial support on which the stochastic model is observed. Consequently, the design of the sampling network influences the amount of information available to estimate the parameters of the geostatistical model and the latent

spatial field introduced in the previous section. In particular, the spatial configuration of the sampled locations affects the estimation of the covariance structure and the accuracy of spatial predictions. Therefore, sampling design and geostatistical modelling should be considered jointly when developing spatial statistical analyses.

2.1 Geostatistical Model for Infractions Detection

The traditional geostatistical formulation can be extended to non-Gaussian outcomes within a generalized linear mixed model framework, as proposed by Diggle *et al.* [20]. In the case of binary responses (as for infraction detection), the observation model assumes that

$$Y_i \sim \text{Bernoulli}(\pi_i), \quad i = 1, \ldots, n, \tag{1}$$

where $\pi_i = P(Y_i = 1 \mid s_i)$ represents the probability of an event (e.g., an alleged infraction) occurring at location s_i.

The Bernoulli success probability is linked to a linear predictor through an appropriate link function. While the logit link is commonly adopted, the complementary log-log (cloglog) link is often preferred when modeling rare events. Under the cloglog specification, the model is written as

$$\text{cloglog} = \log\left(-\log(1-\pi_i)\right) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + u_i, \quad i = 1, \ldots, n, \tag{2}$$

where β_0 denotes the intercept, $\mathbf{x}_i$ is a vector of covariates observed at location s_i, $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients, and $u_i = U(s_i)$ represents a spatial random effect. $\{U(s) : s \in D \subset \mathbb{R}^2\}$ is a latent Gaussian stochastic process describing the underlying spatial phenomenon, with zero-mean and Matérn covariance function defined, for two locations s_i and s_j separated by Euclidean distance $h = \|s_i - s_j\|$, as,

$$C(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\kappa h)^\nu K_\nu(\kappa h), \tag{3}$$

where σ^2 denotes the marginal variance of the spatial field, $\nu > 0$ is a smoothness parameter controlling the differentiability of the process, $\kappa > 0$ is a scale parameter related to the spatial range, and $K_\nu(\cdot)$ denotes the modified Bessel function of the second kind. The Matérn family provides a flexible class of covariance functions capable of representing a wide range of spatial dependence structures [21,22].

This model has a hierarchical representation consisting of the data model $[\boldsymbol{Y} \mid \boldsymbol{w}]$, in (1) and (2), and of the latent spatial structure $[\boldsymbol{w} \mid \boldsymbol{\theta}]$, given by the vector of latent Gaussian effects $\boldsymbol{w} = (\beta_0, \boldsymbol{\beta}, \mathbf{u})$. Vector $\boldsymbol{\theta}$ denotes the set of hyperparameters controlling the covariance structure. Under a Bayesian framework, prior distributions are assigned to all components of $\boldsymbol{w}$ and $\boldsymbol{\theta}$, completing the hierarchical specification through the prior model $[\boldsymbol{w} \mid \boldsymbol{\theta}]\,[\boldsymbol{\theta}]$.

2.2 Adaptive Sampling Design Criteria for Presence/absence Spatial Data

Design-based and model-based approaches are widely used in classical geostatistical problems, particularly when the response variable follows a Gaussian distribution. However, the present study considers presence/absence spatial data, and therefore some sampling design criteria proposed for continuous data may not be directly applicable in this context.

Adaptive geostatistical sampling provides a flexible framework in which new sampling locations are selected sequentially according to the information obtained in previous stages of the study. As discussed by Chipeta et al. [23], Pearse et al. [24], Simões et al. [25], and Silva et al. [26], this strategy allows the data collection process to focus on regions that are expected to provide more informative observations for the objectives of the analysis.

In general, constructing an adaptive geostatistical sampling design for binary data involves three main steps: (i) fitting a geostatistical model to the observed data; (ii) using the fitted model to generate predictions over a grid of potential sampling locations; and (iii) selecting new sampling points according to a predefined selection criterion.

Within this framework, two adaptive sampling criteria are used for binary geostatistical data: the Maximum Risk Sampling Design (MaxRSD) and the Maximum Variance Risk Sampling Design (MaxVRSD), introduced by Miguel, et al. [27].

Maximum Estimated Risk Sampling Design (MaxRSD).

The MaxRSD criterion guides the sampling process towards regions where the estimated probability of occurrence of the phenomenon is larger, meeting the objective of monitoring areas with higher risk of event occurrence. Being the spatial dependence structure initially unknown, a geostatistical model is firstly fitted to the available data and used to predict probability of the event under consideration (risk) on a dense grid of points covering the study region, defining a set of candidate sampling locations. A fixed number of sampling points n is subsequently selected from this grid, assigning to each point a selection probability proportional to the predicted probability of occurrence of the event. As a result, sampling tends to concentrate in areas with higher estimated risk. This criterion depends on the number of sampling points n and on the fitted model m, and is denoted by $\text{MaxRSD}(n, m)$.

Maximum Variance Risk Sampling Design (MaxVRSD).

As an alternative strategy, the MaxVRSD criterion directs the sampling process towards regions where the uncertainty associated with the estimated risk is higher. Similarly to the previous method, a geostatistical model is first fitted to the observed data and risk predictions are obtained over a grid of candidate locations covering the study region. These predictions provide both the estimated probabilities of event occurrence and their associated standard errors. New sampling locations are then selected with probability proportional to the magnitude of the standard errors of the predicted probabilities, thereby concentrating data

collection in areas where the model presents greater uncertainty. This criterion also depends on the number of sampling points n and the fitted model m, and is denoted by MaxVRSD(n, m).

For comparison purposes, a Stratified Random Sampling (SRS) design is considered as a reference method. This sampling strategy is widely used in fisheries-related studies and is well supported by classical sampling theory [28]. In a spatial context, this approach divides the study region into several sub-regions or strata that are relatively homogeneous with respect to the variable of interest. A sampling location is then randomly selected within each stratum, ensuring that all parts of the study area are represented in the sample and providing balanced spatial coverage.

In this study, stratified random sampling is used as a benchmark to evaluate the performance of the proposed adaptive criteria, namely MaxRSD and MaxVRSD. Unlike these adaptive approaches, which select new sampling locations based on the estimated risk or the uncertainty of model predictions, the stratified design follows a fixed strategy based solely on the spatial partition of the study area.

3 Sampling Design Performance Evaluation

The performance of the different sampling designs considered in this study is assessed based on the predictive quality of the geostatistical models fitted to the data. For this purpose, three commonly used predictive performance measures are considered: the Root Mean Squared Error (RMSE), the Logarithmic Score (Log-Score), and the Leave-Group-Out Cross-Validation (LGOCV). These metrics allow different aspects of predictive performance to be evaluated, namely the accuracy of point predictions, the quality of the predictive distribution, and the model's generalization ability.

3.1 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a widely used measure to quantify the discrepancy between observed values and model predictions. This metric provides an indication of prediction accuracy, assigning greater penalties to larger errors due to the squaring of the differences. Formally, RMSE is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{4}$$

where y_i denotes the observed value at location s_i, $\hat{y}_i$ represents the predicted value obtained from the geostatistical model for the same location, and n is the total number of observations considered. Lower RMSE values indicate better predictive performance [29,30].

3.2 Logarithmic Score (Log-Score)

While RMSE evaluates only the distance between observed values and point predictions, the Logarithmic Score (Log-Score) assesses the quality of the full predictive distribution associated with the model. This metric is based on the probability assigned by the model to the observed values. The Log-Score is defined as

$$\text{Log-Score} = -\frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid \hat{\theta}), \tag{5}$$

where $f(y_i \mid \hat{\theta})$ denotes the predictive probability density (or probability mass) of the observed value y_i under the model estimated with parameters $\hat{\theta}$. Smaller values of the Log-Score indicate better predictive performance, since they imply that the model assigns higher probability to the observed data [9,31].

3.3 Leave-Group-Out Cross-Validation (LGOCV)

Leave-Group-Out Cross-Validation (LGOCV) is used to evaluate the generalization capability of the model. In this procedure, the dataset is divided into K groups. At each iteration, one group is removed from the training data, the model is fitted to the remaining data, and predictions are then generated for the excluded observations. The overall LGOCV error can be expressed as

$$\text{Error}_{LGOCV} = \frac{1}{K}\sum_{k=1}^{K} L\left(y^{(k)}, \hat{y}^{(-k)}\right), \tag{6}$$

where $y^{(k)}$ represents the observations belonging to group k, $\hat{y}^{(-k)}$ denotes the predictions obtained from the model fitted without that group, and $L(\cdot)$ represents an appropriate loss function, such as RMSE or Log-Score [32].

This procedure provides a robust way to assess predictive performance, reducing the risk of overfitting and offering a more realistic estimate of the model's ability to predict new observations.

4 Maritime Fishery Surveillance at Central of Portugal

The Portuguese naval deployment is characterised by operational flexibility, with no permanent public allocation of assets to a specific geographical area. Resources are distributed according to strategic priorities, the intensity of maritime activity, and enforcement requirements. In the central region of Portugal, extending between Figueira da Foz, Lisbon, and Sines, maritime surveillance is ensured by different types of vessels operated by the Portuguese Navy and by units of the National Maritime Authority [2].

Within the Navy, Coastal Patrol Vessels (CPVs) include vessels such as NRP[1] Zaire, NRP Tejo, NRP Douro, and NRP Mondego (See the Fig. 1). These smaller

[1] Portuguese Republic Ships - NRP

and highly manoeuvrable vessels are particularly suited to missions within the territorial sea and coastal waters. Their operational role focuses on fisheries inspection, routine boarding operations, maritime traffic control, and reinforcing the State's presence in areas with a high concentration of vessels [33].

In addition, Offshore Patrol Vessels (OPVs) of the Viana do Castelo class, namely NRP Sines and NRP Figueira da Foz, operate in the region. These vessels have greater autonomy, extended endurance at sea, and the capability to operate both in the territorial sea and across the Exclusive Economic Zone (EEZ). Their missions include large-scale surveillance, offshore inspection and control, support to search and rescue (SAR) operations, environmental monitoring, and interinstitutional cooperation. Their enhanced operational capacity allows coverage of extensive areas of the central coastline and adjacent offshore waters [33].

From a spatial standpoint, patrol resources are distributed according to their operational roles. Coastal patrol units and Maritime Police vessels are typically deployed near the shoreline and around major ports, whereas offshore patrol vessels cover wider maritime areas, including distant offshore zones. Consequently, the spatial distribution of enforcement operations and recorded infringements may reflect both the level of maritime activity and the surveillance strategy adopted. Within this framework, factors such as distance from the coastline, proximity to ports, and the spatial extent of the Exclusive Economic Zone (EEZ) can act as structural drivers of the variability observed in enforcement data.



Fig. 1. Vessels of the naval deployment

In this section data are presented and modelled, allowing the sampling design methods to be applied.

4.1 Data

The dataset used in this study consists of georeferenced records of inspection activities carried out by the Portuguese Navy along the central region of the Portuguese coast over a five-year period (2013−2017). Inspection operations were organized into four daily periods of six hours each: 00:01−06:00, 06:01−12:00, 12:01−18:00, and 18:01−24:00.

The analysis focuses on three types of fishing-related infractions, recorded in binary form (presence/absence): the use of prohibited fishing gear (infraction 3),

fishing in restricted areas (infraction 4), and fishing prohibited due to excessive engine power or vessel tonnage (infraction 5).

Figure 2 illustrates the spatial distribution of inspection locations recorded during the study period. Red points indicate locations where suspected infractions were detected, while black points represent locations where no infractions were identified.

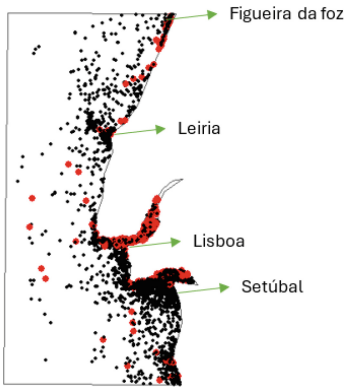


Fig. 2. Maritime data distribution map of the Figueira da Foz, Lisbon, and Sines coastal region

4.2 Geostatistical Model for Presence/absence Data

In this study, the available data correspond to binary observations indicating the presence or absence of alleged fishing infractions at specific locations [10]. As discussed in Sect. 2.1, these can be described by the Geostatistical model for binary data in (1), (2) and (3).

A total of 19 models were fitted, as presented in Table 1, with the aim of comparing different model specifications. The models were fitted efficiently using R-INLA and compared using the Deviance Information Criteria, DIC. Among these, Model O stood out by achieving the lowest DIC value (5687.46), indicating the best balance between model fit and complexity. This model incorporates distance, year, and period as covariates, along with a spatial random effect, allowing for an effective representation of the spatial variability of the phenomenon under study. Consequently, this model was selected as the most appropriate, and its results are presented next.

The estimation results for the fixed effects associated with the covariates distance to the coast, year, and period in Model O are presented in Table 2, together with the estimated hyperparameters. The variance explained by the Matérn spatial field (σ_u^2) was estimated at 3.725, while the posterior mean of the range parameter (R) was 8.477 km, with a 95% credibility interval of (5.616,

Table 1. Fitted models and corresponding DIC

Model	Linear Predictor	DIC
A	$\beta_0 + \boldsymbol{u}$	6593.49
B	$\beta_0 +$ period $+ \boldsymbol{u}$	6524.66
C	$\beta_0 +$ year $+ \boldsymbol{u}$	6551.73
D	$\beta_0 +$ year + period $+ \boldsymbol{u}$	6482.10
E	$\beta_0 +$ bathymetry + year + period $+ \boldsymbol{u}$	6543.86
F	$\beta_0 +$ bathymetry $+ \boldsymbol{u}$	6646.52
G	$\beta_0 +$ bathymetry + year $+ \boldsymbol{u}$	6604.05
H	$\beta_0 +$ bathymetry + period $+ \boldsymbol{u}$	6584.34
I	$\beta_0 +$ bathymetry + year + period	7970.26
J	$\beta_0 +$ bathymetry + year	8032.84
K	$\beta_0 +$ bathymetry + period	8017.94
L	$\beta_0 +$ bathymetry	8076.14
M	$\beta_0 +$ bathymetry + slope + distance + year + period $+ \boldsymbol{u}$	5689.46
N	$\beta_0 +$ slope + distance + year + period $+ \boldsymbol{u}$	5688.88
O	$\beta_0 +$ distance + year + period $+ \boldsymbol{u}$	**5687.46**
P	$\beta_0 +$ distance + period $+ \boldsymbol{u}$	5723.40
Q	$\beta_0 +$ distance + year $+ \boldsymbol{u}$	5738.05
R	$\beta_0 +$ distance $+ \boldsymbol{u}$	5772.27
S	$\beta_0 +$ year + period $+ \boldsymbol{u}$	5691.35

12.235). The effects of the working periods of the day, as well as all years from 2013 to 2017, were found to be statistically significant.

Figure 3 presents the estimated risk surfaces associated with fishing-related infractions for the year 2014. Overall, a clear spatial pattern is observed, with higher-risk areas concentrated along the coastal strip. Spatial dependence is pronounced, as reflected in the smooth transitions between low- and high-probability zones. During nighttime periods, shown in panels (a) and (b) for the year of 2014, higher-intensity areas are particularly evident in the northern (Leiria) and central (Lisbon) coastal regions, appearing slightly more extensive than in previous years, which suggests an expansion of the risk area. In contrast, daytime periods, represented in panels (c) and (d) for the year of 2014, exhibit a more diffuse spatial distribution, although localized increases in probability are observed in areas coinciding with nighttime hotspots. This indicates that certain high-risk locations remain consistently active across different periods of the day. The estimated risk surfaces associated with presumed fishing infractions for the years of 2013, 2015, 2016 and 2017 are consistent with that of year 2014 just described, only with a somewhat decreased density of occurrences for the year of 2013.

Considering the overall study period (2013−2017), the results reinforce the existence of a consistent and persistent spatial pattern, with recurring hotspots

Table 2. Summary of the parameter's posterior distributions of the model O

Variable	Mean	SD	Median	95% Credibility Interval
(Intercept)	−1.715	0.193	−1.705	(−2.124, −1.365)
dist_coast_std	−0.225	0.158	−0.226	(−0.535, 0.086)
Year 2014	0.389	0.106	0.389	(0.180, 0.597)
Year 2015	0.521	0.117	0.521	(0.292, 0.750)
Year 2016	0.755	0.120	0.755	(0.519, 0.992)
Year 2017	0.266	0.134	0.266	(0.004, 0.529)
06:01–12:00	−1.122	0.157	−1.122	(−1.431, −0.813)
12:01–18:00	−1.083	0.161	−1.083	(−1.400, −0.767)
18:01–24:00	−0.660	0.186	−0.660	(−1.024, −0.296)
σ_u^2	3.725	0.790	3.637	(2.422, 5.515)
R	8.477	1.691	8.319	(5.616, 12.235)

along the coastline. Risk levels are systematically higher and more concentrated during nighttime periods, while daytime patterns tend to be more diffuse and less intense.

4.3 Sampling Design

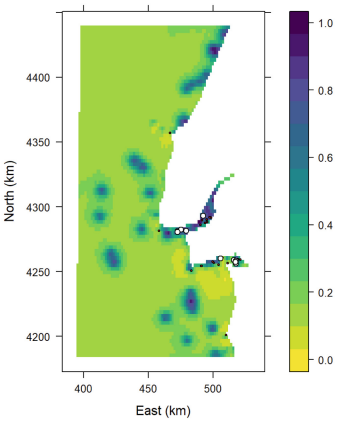
After obtaining the fitted value maps of fishing-related infractions, the same model is used to construct the sampling designs. In this context, the MaxRSD and MaxVRSD designs proposed by Miguel et al. [27] are considered and compared with the SRS design, assuming a success probability fixed at $p = 0.2$.

Figure 6 in the appendix, compares the SRS, MaxRSD and MaxVRSD sampling designs for 50 points, highlighting distinct spatial patterns. SRS provides a random coverage, MaxRSD focuses on high-risk areas, and MaxVRSD balances risk and uncertainty, supporting improved model estimation.

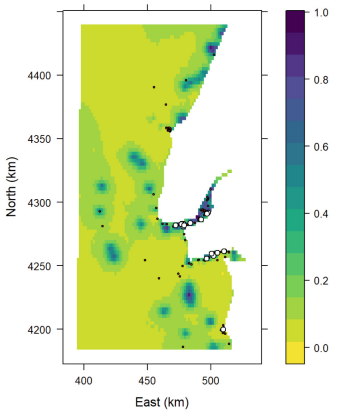
Figure 4 shows the three sampling designs (SRS, MaxRSD and MaxVRSD) for 200 points, where the differences between them become less evident. SRS maintains a more random distribution, MaxRSD continues to focus on high-risk areas, and MaxVRSD provides a balance between risk and uncertainty.

4.4 Surveillance Routes

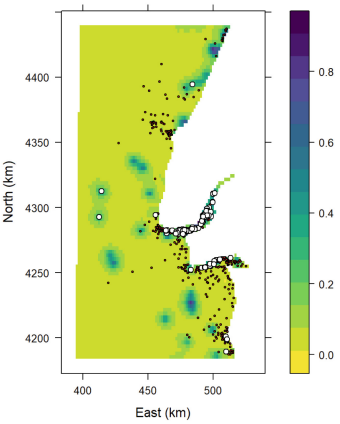
The resulting risk maps highlight areas where infractions are more likely to occur, thereby assisting inspection planning, which combined with the proposed sampling designs, allow to improve the allocation of patrol resources, potentially reducing operational costs, providing support for the Portuguese Navy's maritime surveillance activities related to fisheries inspection. Figure 5 (a) and (b) illustrate two distinct operational strategies based on the MaxRSD sampling design for 200 points, highlighting how route organisation can influence the efficiency of surveillance activities.



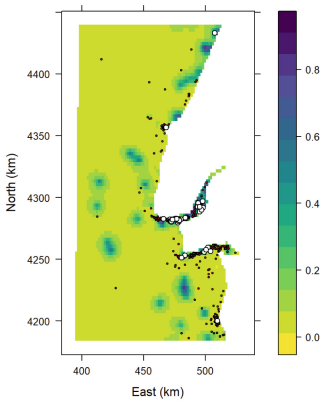
(a) Night: 0:01–6:00



(b) Night: 18:01–24:00



(c) Day: 06:01–12:00

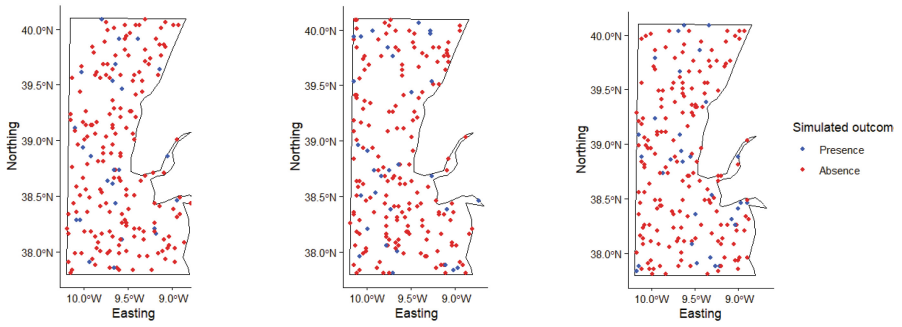


(d) Day: 12:01–18:00

Fig. 3. Map of fitted values for the 1st period (top-left) and 4th period (top-right), and below the 2nd period (bottom-left) and 3rd period (bottom-right) for the year 2014.

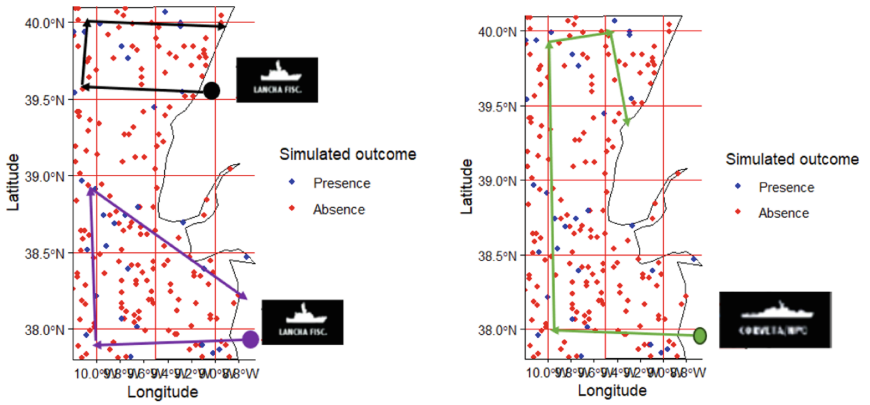
In Fig. 5 (a), two independent routes are proposed, starting from Setúbal and Leiria. This configuration enables simultaneous operations in different areas, particularly in zones with a higher concentration of occurrences, thereby improving response time and optimising travel efficiency.

In contrast, Fig. 5 (b) presents a single continuous route extending from Setúbal to Figueira da Foz. Although this approach may involve a longer trajectory, it provides a more uniform spatial coverage and allows for a more integrated monitoring of the study area along the coastline.



(a) SRS (b) MaxRSD (c) MaxVRSD

Fig. 4. Comparison of sampling designs (200 simulated points, Year 2014).



(a) MaxRSD sampling maps (200 points) with two proposed enforcement routes.

(b) MaxRSD sampling maps (200 points) with one proposed enforcement routes.

Fig. 5. Proposed surveillance routes

4.5 Performance Design Evaluation

Table 3 presents the comparison of the SRS, MaxRSD, and MaxVRSD sampling designs using different predictive performance criteria. The results show that the effectiveness of each sampling strategy strongly depends on sample size and on the specific objective of the analysis.

Under sparse sampling conditions (50 points), substantial discrepancies were observed among the models, particularly in the LGOCV values, which indicate reduced predictive stability and increased uncertainty when limited spatial information is available. In this scenario, SRS provides greater overall stability but shows limited ability to identify high-risk areas and capture complex spatial patterns. In contrast, the adaptive designs, MaxRSD and MaxVRSD, more effec-

Table 3. Comparison of sampling design criteria for different sample sizes

Criteria	50 points				200 points			
	RMSE	LogScore	LGOCV	DIC	RMSE	LogScore	LGOCV	DIC
MaxVRSD	0.6051	−6.4451	616.4678	326.7230	0.0429	−0.1572	0.4458	174.4539
MaxRSD	0.2136	−0.3118	705.8629	258.8679	0.3163	−0.5850	0.3749	145.2507
SRS	0.3341	−1.2825	708.0625	91.8115	0.1332	−0.2879	0.5162	156.3470

tively direct sampling towards relevant locations, improving the representation of the spatial structure of the phenomenon.

Among the adaptive approaches, MaxRSD achieved the best balance between predictive accuracy and probabilistic calibration, presenting lower RMSE and better LogScore values. This suggests that the design is more effective for identifying and predicting high-risk areas under constrained sampling conditions. Conversely, MaxVRSD demonstrated a stronger capacity for uncertainty exploration, although at the expense of reduced local prediction precision, as reflected by its higher RMSE and lower LogScore values.

As the sample size increased to 200 points, all models exhibited substantial improvements in predictive stability, with LGOCV values decreasing considerably. In this case, the differences between the sampling strategies became less pronounced, although the model-based designs continued to show superior overall performance. MaxVRSD achieved the highest predictive accuracy, while MaxRSD maintained a strong ability to identify critical areas. Although SRS also improved with denser spatial coverage, it remained less efficient in highly heterogeneous spatial contexts.

Overall, the results suggest that adaptive sampling designs based on risk and predictive uncertainty provide a more efficient framework for optimising surveillance resources and improving enforcement strategies, particularly in operational scenarios characterised by limited sampling capacity.

4.6 Conclusions and Recommendations

Optimising enforcement efforts remains a significant challenge for maritime authorities, and sampling design plays a key role in improving resource management and reducing operational costs. The results of this study show that the choice of sampling design has a substantial impact on both the quality of spatial inference and the effectiveness of field operations.

The comparison between SRS, MaxRSD and MaxVRSD demonstrates the advantages of model-based approaches. For smaller sample sizes, SRS provides balanced spatial coverage but shows limitations in identifying critical areas. In contrast, MaxRSD directs sampling towards regions with higher probabilities of infractions, enhancing detection capability, while MaxVRSD incorporates both risk and uncertainty, leading to improved estimation quality.

As the sample size increases, the differences between the sampling designs become less pronounced. Nevertheless, approaches that incorporate predicted probability and variability continue to offer advantages, particularly in resource-constrained settings. Furthermore, the integration of sampling design with operational route planning proves to be crucial, as the spatial organisation of enforcement activities can significantly affect overall efficiency.

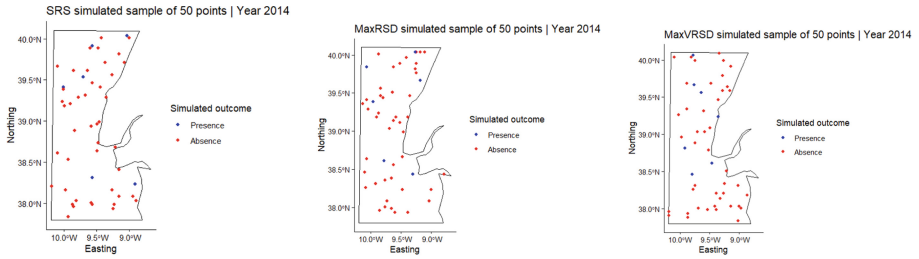
Overall, the findings confirm that model-based sampling designs are effective tools for supporting fisheries enforcement. MaxRSD stands out for its ability to prioritise high-risk areas, while MaxVRSD is particularly useful for improving estimation by accounting for uncertainty. Although SRS is simple and widely used, it shows limitations in contexts with high spatial heterogeneity. In this sense, combining risk maps with adaptive sampling strategies enables more efficient use of resources and supports better-informed operational decisions.

Based on these findings, the following recommendations are proposed: adopt MaxRSD when the primary objective is to maximise detection of infractions; use MaxVRSD to improve estimation and reduce uncertainty; integrate sampling design with operational route planning; develop patrol routes tailored to specific types of infractions; and explore simulation-based approaches and decision-support tools to optimise resource allocation and enforcement strategies. With regard to integrate sampling design with operational route planning, it is expected that in the near future, along with the development of appropriate software will be able to simulate patrol routes allowing the selection of the target zone for simulate a feasible patrol route depending on the infraction, or set of infractions, takingo into accout available the operational resources, to achive the highest probability of success for missions carried out by the Portuguese maritime surveillance.

Acknowledgments. This work is funded by national funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the projects UID/00297/2025 (https://doi.org/10.54499/UID/00297/2025) and UID/PRR/00297/2025 (https://doi.org/10.54499/UID/PRR/00297/2025) (Center for Mathematics and Applications). We thank the NOVA Math and the Portuguese Navy for making this work possible, within the scope of the protocol established between them.

Disclosure of Interest. The authors have no competing interests to declare that are relevant to the content of this article.

A Appendix



(a) SRS (b) MaxRSD (c) MaxVRSD

Fig. 6. Comparison of sampling designs (50 simulated points, Year 2014).

References

1. Portuguesa, M.: Navios Patrulha Oceânicos – Classe Viana do Castelo (2023). https://www.marinha.pt
2. Autoridade Marítima Nacional. Missão e Competências da Polícia Marítima (2023). https://www.amn.pt
3. Guarda Nacional Republicana. Unidade de Controlo Costeiro – Missão (2023). https://www.gnr.pt
4. Bueger, C., Edmunds, T.: Beyond seablindness: a new agenda for maritime security studies. Int. Aff. **93**(6), 1293–1311 (2017)
5. European Commission. EU Maritime Security Strategy Action Plan. European Commission (2021)
6. Jr, H., David, W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 398. John Wiley & Sons, New York (2013)
7. McCulloch, C.E., Searle, S.R.: Generalized, Linear and Mixed Models. John Wiley & Sons, New York (2004)
8. De Oliveira, V.: Models for Geostatistical Binary Data: Properties and Connections. The American Statistician, vol. 74, no. 1, pp. 72–79. Taylor & Francis (2020)
9. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. J. Royal Stat. Soc. Seri. b (Stat. Methodol.) **71**(2), 319–392 (2009). Wiley Online Library
10. Miguel, B., Simões, P., de Deus, R. G., and Natário, I.: Bayesian geostatistics modeling of maritime surveillance data. In: Proceedings of the International Conference on Computational Science and Its Applications, pp. 187–205. Springer (2024)
11. Petitgas, P.: Geostatistics in fisheries survey design and stock assessment: models, variances and applications. Fish Fish. **2**(3), 231–249 (2001)
12. Brus, D.J., De Gruijter, J.J.: Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma **80**(1–2), 1–44 (1997)

13. De Gruijter, J.J., Ter Braak, C.J.F.: Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Math. Geol. **22**(4), 407–415 (1990)
14. Liu, Y., Chen, Y., Cheng, J.: A comparative study of optimization methods and conventional methods for sampling design in fishery-independent surveys. ICES J. Mar. Sci. **66**(9), 1873–1882 (2009)
15. Diggle, P., Lophaven, S.: Bayesian geostatistical design. Scand. J. Stat. **33**(1), 53–64 (2006)
16. Van Groenigen, J.W., Siderius, W., Stein, A.: Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma **87**(3–4), 239–259 (1999)
17. Russo, D.: Design of an optimal sampling network for estimating the variogram. Soil Sci. Soc. Am. J. **48**(4), 708–716 (1984)
18. Simbahan, G.C., Dobermann, A.: Sampling optimization based on secondary information and its utilization in soil carbon mapping. Geoderma **133**(3–4), 345–362 (2006)
19. Brus, D.J.: Spatial Sampling with R. CRC Press, Boca Raton (2022)
20. Peter, J.D., Jonathan, A.T., Rana, M.A.: Model-based geostatistics. J. Royal Stat. Soc. Seri. C (Appl. Stat.) **47**(3), 299–350 (1998). Wiley Online Library
21. Finn, L., Håvard, R., Johan, L.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. Royal Stat. Soc. Seri. B (Stat. Methodol.) **73**(4), 423–498 (2011). Wiley Online Library
22. Elias, K., et al.: Advanced Spatial Modeling With Stochastic Partial Differential Equations Using R and INLA. Chapman and Hall/CRC (2018)
23. Chipeta, M.G., Rowlingson, B., Diggle, P.J.: Geosample: an R Package for Geostatistical Sampling Designs (2019). Under review
24. Pearse, A.R., et al.: SSNdesign: an r package for pseudo-Bayesian optimal and adaptive sampling designs on stream networks. PLoS ONE **15**(9), e0238422 (2020)
25. Simões, P., Carvalho, M.L., Figueiredo, I., Monteiro, A., Natário, I.: Geostatistical sampling designs under preferential sampling for black scabbardfish. In: Proceedings of the International Conference of the Portuguese Statistical Society, pp. 137–151. Springer (2021)
26. Silva, D., Menezes, R., Serra-Pereira, B., Azevedo, M., Figueiredo, I.: Adapting the sampling design of research surveys to improve the biomass estimation of secondary species—the case study of Raja clavata. ICES J. Mar. Sci. **80**(10), 2567–2578 (2023)
27. Miguel, B., Simões, P., de Deus, R.G., Natário, I.: Sampling design for binary geostatistical data, application to inspection actions of fishing activity in Portugal. Spatial Stat. 100919 (2025)
28. Cochran, W.G.: Sampling Techniques. John Wiley & Sons (1977)
29. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2009)
30. Diggle, P.J., Ribeiro, P.J.: Model-Based Geostatistics. Springer, New York (2007)
31. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. **102**(477), 359–378 (2007)
32. Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. **27**, 1413–1432 (2017)
33. Portuguese Navy. Patrol Vessels – NRP ZAIRE, NRP TEJO, NRP DOURO, NRP MONDEGO. Disponível em: https://www.marinha.pt/en/os_meios/patrulhas/Pages/default.aspx Acesso em 2026

Investigating of the Operation of a Railgun System with Laboratory Proportion

P. Sampaio Pereira[1], Mário Gatta[1,2], and M. Filomena Teodoro[1,3(✉)]

[1] CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, 2810-001 Almada, Portugal
maria.teodoro@tecnico.ulisboa.pt

[2] Center of Philosophy of Sciences, Faculdade de Ciências, University of Lisbon, Lisbon 1749-016, Portugal

[3] CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, 1048-001 Lisboa, Portugal
mteodoro64@gmail.com

Abstract. In this document it is addressed the basic principles underlying the function of the railgun focusing in the electrodynamic questions. It is intended to acquire familiarity with this kind of weapon through the critical study of his functional principles. At first glance, the basic principle of operation of a railgun may seem simple in the light of the immediate application of the Lorenz force in a conductor driven by current. However, everything becomes more complicated in the case of a quick change of the involved parameters (transitional regime), which requires a more thorough analysis of current behavior, electric and magnetic fields, and all involved materials in the system.

This project also included the construction of two railguns, initially a smaller size one to gain familiarity with the system, and lastly a laboratory proportion one in which was performed and analyzed a number of tests with different materials and distinct projectile sizes. To better understand the process that involve the railgun process, an analysis in the time domain, of the spatial distribution of the electromagnetic field, electrical current and energy flow was performed. In the present paper, we describe the preliminary concepts about the studied weapon, the energy flow, complemented by a summary of the experimental work associated to the construction of the second railgun (in laboratory proportions) that allowed us to test distinct materials, forms and sizes.

Keywords: railgun · electric field · magnetic field

1 Introduction

The Lorentz force has been known for over two hundred years, but its application in a railgun was not immediate, and there were differing opinions on the subject. Only recently has its use been considered of great interest, and this interest has increased further with the construction of fully electric warships.

Since the last two decades of last century, the importance of magnetic acceleration weapons (MAGs) such as railgun has grown considerably. Applications are envisioned

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 69–86, 2027.
https://doi.org/10.1007/978-3-032-30521-3_5

to replace traditional rapid acceleration systems in aircraft carriers, launching space platforms, welding, surface coating, and firing projectiles at hyper-velocity as a weapon system. The US Department of Defense, has been invested in the development of this technology greatly, e.g., in projects such as CHECMATE (Combat High Energy Capacitor Module Advanced Technology Experiment), whose main objective is the launch of ballistic missiles through a ground-based system.

With the end of the Cold War, in 1990, the enthusiasm surrounding this research reduced. With the 21st century and the fully electric ships of the U.S. Navy, the interest to lead to better-performing railguns to its new generation reborned. [1].

Railguns are almost as old as the concept of the coil gun (Gauss cannon or electromagnetic cannon). These accelerate a conductive electrical design that bridges a pair of current-carrying rails. The advantages that this technology brings to weapons systems include very high speeds (2 km/s or 61 $Mach$[1]), propellant-free ammunition, designs are much cheaper per firing and less bulky, greater safety in ammunition storage and the ability to use primary energy sources (electrical in tanks and ships) in a "parasitic" way instead of chemical expansion of gases that are toxic to humans.

Currently, the U.S. Navy is investing in the ultimate goal of achieving the capability of a weapon, with *mega joule* energy items, capable of making continuous fire at speeds exceeding 6 *Mach*, exceeding 200 nautical *miles* of range [1].

It is the interest of Portuguese Navy to pursue this technological evolution, evaluating the advantages of the adoption of this kind of weapon. In this document it is addressed the basic principles underlying the function of the railgun focusing in the electrodynamic questions. It is intended to acquire familiarity with this new kind of weapon through the critical study of his functional principles.

2 Basic Operation of a Railgun

It is currently widely accepted that a railgun projectile is accelerated by the Lorentz force. This force is described by equation (1) which describes the force on a particle with charge q moving with velocity v in the presence of an electric field E and a magnetic field B.

$$F = q\left[E + (v \times B)\right] \tag{1}$$

In the case of a railgun, the interacting field is magnetic, so it is possible to simply rewrite Equation (1) into (2) .

$$F = q\,(v \times B)\,. \tag{2}$$

The force element is written as $dF = idl \times B$ over a elemtent of flow idl.

2.1 Structure

The basic construction of a railgun is quite simple (see the Fig. 1), it only requires:

[1] 1 *Mach* is defined as the ratio between the speed of the object and the speed of sound.

- 2 metal conductors (which will serve as rails);
- 1 metal conductor that bridges the two conductors (the projectile itself or a projectile support (armor));
- 1 power supply (preferably direct current);
- 1 switch

Both types of current (direct and alternating) will accelerate the railgun projectile; however, pulsed current (direct or DC) is generally used for more powerful shots [2].

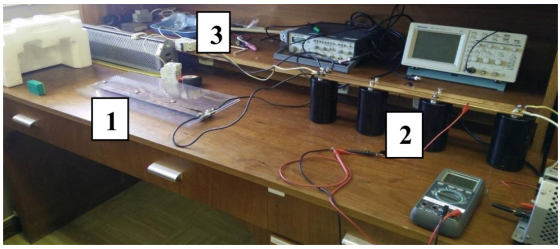


Fig. 1. Assembly of the second railgun built in this study. Legend: 1 – Parallel rails; 2 – Capacitor bank (Power Supply); 3 – Switch.

The architecture of a railgun consists of two metallic conductors that serve as rails, and an electrically conductive projectile that closes the circuit, thus connecting the two rails. The current flows through one rail, passes through the projectile, and returns to the source through the other rail, creating a magnetic field (see the Fig. 2).

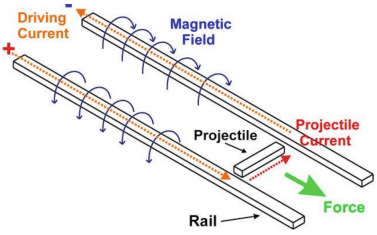


Fig. 2. Diagram of the interaction between the magnetic field B and the current flowing in the projectile. Source [3].

The magnetic field generates a Lorentz force, pointing outwards, in all components of this current loop. Railguns take advantage of this phenomenon to accelerate the projectile out of the rails.

The Lorentz force equation in its original form can be complex to work with in practical cases like railguns. To simplify these cases, equation (2) was rewritten as a function of the current I and the inductance gradient L'. The derivation of equation (3) from equation (2) and the calculation of L' from the rail geometry can be found in [1] and it is given by

$$F = \frac{L' I^2}{2}. \tag{3}$$

The inductance gradient is given by Eq. (4):

$$L' = \frac{\mu_0 I^2}{2\pi} ln\left(\frac{d-r}{r} Y\right). \tag{4}$$

where d represents the length of the rails and r the distance between them.

3 Model

The assembly of a conventional railgun is simple. The rails must be parallel and the projectile in the middle of the rails and in contact with them, so that the projectile closes the electrical contact, as shown in Fig. 3, on the right.



Fig. 3. Left - prototype tested on the destroyer ship "Zumwalt" (US Navy). Right, scheme of a conventional railgun that uses a sabot to close the electrical contact.

A high-voltage power supply (typically banks of high-voltage and high-capacitance capacitors) is needed to rapidly discharge the circuit, accelerating the projectile.

The objects that railguns fire are called projectiles or armatures.

When the goal is to hit the target with the complete object that was fired, it is called a projectile; if a sabot is used to close the electrical contacts, the assembly (projectile and sabot) is called an armature. Currently, in railgun weapon systems, a conductive sabot is used to make contact with the rails. As soon as the assembly leaves the rails, the sabot is separated from the projectile due to its less aerodynamic shape.

4 Problems in Railgun System

The fact that certain Navies are ready to use this type of weaponry means that railguns, which were once unreliable and had a short lifespan, have overcome or at least reduced the problem of rail damage. This damage can take many forms due to the extreme conditions to which they are exposed.

High temperatures, metal-to-metal friction, high projectile speeds are just some of the conditions that the rails will have to withstand for the longevity of a railgun to be viable for military use.

4.1 Excavation or Gouging

Excavation occurs due to high sliding speeds, as happens in missile and railgun tracks. This effect takes the form of craters similar to a drop of water (see Fig. 4, top), and occurs above a limiting speed value depending on the friction and properties of the involved materials [4]. Excavation is a problem for two reasons: on the one hand, this process generates a very high dynamic stress that can lead to failures in both the rails and the projectile of a railgun; on the other hand, the use of material already with this effect can lead to areas of non-contact with the armor. Most importantly, when excavation occurs, a tremendous force is exerted on the armor in the opposite direction of movement, substantially reducing the performance of the railgun. Excavation became a problem in railguns when they began to be built with solid armor accelerators in 1980. In the construction of a railgun, the most important criterion in choosing the material for rails is their high electrical conductivity. Copper for rails and aluminum for armor are the most common choices. Replacing the rails with aluminum and the projectile with copper would substantially reduce this effect, however, the yield would inevitably drop because aluminum is a worse conductor than copper. A copper projectile is not the best choice because it is a denser material than aluminum, making it heavier [5].



Fig. 4. Left: Excavation in copper rails. The black arrow indicates the direction of movement. Source [4]. Center: Damage inflicted by the bows on the railgun built by Power Labs. Source [6]. Right: Example of an aluminum tank verified on the Railgun *MK*1.1. Source [7].

4.2 Current Arcs or Arcing

Generally, arcs occur when the contacts between the rails and the projectile are not perfect. During the projectile's path along the rails, it has been recorded that half of its mass is lost due to heating and friction at high speeds ([5]. When there is a small gap between the projectile and the rails, and if the electromagnetic field is strong enough, arcs are formed and the current travels through the air until it reaches the projectile. These current jumps through the air decrease the efficiency of the railgun because the current travels through areas of very high resistance (air), losing a large part of its energy in the formation of current arcs. The damage caused by the arcs is visible and deteriorating to the rails, as can be seen in Fig. 4, on left.

4.3 Deposits Excess

Deposition occurs during projectile firing when it heats up due to high speeds and friction, leaving part of its mass in a liquid state on the rails (see Fig. 4, on right). This layer of liquid aluminum is beneficial because it acts as a lubricant during firing [5]. Subsequently, these aluminum deposits cool and solidify, becoming very irregular, flaky, and porous, reducing the performance of subsequent shots. Accuracy in the distance between rails is crucial in a railgun because any disturbance

in the caliber results in loss of electrical contact, less efficiency, and decreased longevity of the weapon. Therefore, one way to reduce the amount of deposits is to improve the electrical contact between the rails and the projectile, so that there is a uniform zone with compatible sections, avoiding as much as possible areas of constriction for the current.

5 Power Supplies

The low efficiency level of a railgun is practically inevitable. Even using techniques to improve the weapon's performance, such as the use of electrically conductive lubricant, low electrical resistance rails, and projectile pre-acceleration, we still have efficiencies below 40%, and in the best cases it is possible to reach up to 66% [8,9].

Due to the low yields, to achieve speeds greater than Mach 6, such as the new weapon developed by the US Navy [10], enormous amounts of stored energy (on the order of MJ) are needed, which are subsequently released almost instantaneously in the projectile. Typically, capacitor banks are used to store electrical energy. To release all the stored energy, switching mechanisms are used that can handle high currents. The most common components are thyristors or spark gap switches. The spark gap switch are less common. They consist of an assembly with two conductive electrodes separated by a space filled with air or gas. This is designed to allow an electric spark to pass between the conductors when the voltage difference reaches the breakdown voltage of the gas in that space [11]. Table 1 compares different railgun configurations already tested, as well as the voltage of the capacitors and activation mode of the projectile [6,12,13].

Table 1. Scale of stored energy in different railgun designs.

Railgun	Number of capacitors	Stored Energy	Activation Mode
1º Proj	3 (385 V)	500J	None
2º Proj	13 (400 V)	11 kJ	None
Powerlabs 2.0	32 (450 V)	20 kJ	None
Cal Poly Mk 1.0	16 (450 V)	16 kJ	None
Cal Poly Mk1.1	16 (450 V)	16 kJ	Thyristors
Cal Poly Mk 2.0	2 (9 kV)	67 kJ	*Spark Gap Switch*
GA Blitzer	Unknown	2 MJ	Solid-State Switch
GA/US *Navy*	>200 (11kV)	9 MJ	Solid-State Switch
BAE//*Navy*	29 bench (11kV)	>32 MJ	Unknown

6 Charging and Discharging Capacitors

6.1 Charging Phase

The energy required for a railgun to operate is stored in the capacitors, in the form of electrical energy in the field between the plates. It is important to know, rigorously, their behavior during the discharge phase, because in addition to the typical capacitor discharge curve, there are other variables depending on time, such as electrical resistance and Lorentz force.

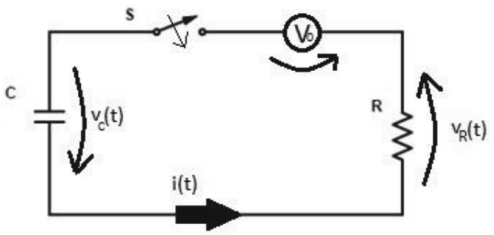


Fig. 5. Diagram of an RC circuit.

The charging circuit of a railgun can be approached as a circuit that consists of a resistor **R**, a capacitor**C**, and a source **V**, in series. This circuit is known as an **RC** circuit, as shown in the Fig. 5.

The charging of a capacitor in an *RC* circuit is described by the voltage across its terminals, $v_c(t)$, and $i(t)$, the current flowing through the circuit.

$$v_c(t) = V_o\left(1 - e^{-\frac{t}{RC}}\right), \tag{5}$$

$$i(t) = I_o e^{-t/RC}. \tag{6}$$

$$\left| \begin{array}{ll} v_R(t) = Ri(t) = R\dfrac{dq_c(t)}{dt}, & (7) \\ v_c(t) = \dfrac{1}{C}q_c(t), & (8) \end{array} \right.$$

$$-v_0 + v_R(t) + v_c(t) = 0 \rightarrow \tag{9}$$

$$\rightarrow -v_0 + R\frac{dq(t)}{dt} + \frac{1}{C}q(t) = 0 \rightarrow \tag{10}$$

$$\rightarrow \frac{dq(t)}{dt} + \frac{1}{RC}q(t) = \frac{1}{R}v_0, \tag{11}$$

Equation (5) models the voltage across the capacitor as a function of time after the switch *s* is closed, $v_c(t)$ is the source voltage, *R* is the resistance value in the circuit, and *C* is the capacitance of the capacitor *C*. Equation (6) represents the current $I(t)$ as a function of time *t* and *I(0)* the peak current, with $I(0) = \frac{V_0}{R}$. These equations can be obtained by solving the differential equation of an *RC* circuit under load, as we know from electrical engineering, $q_c(t)$ represents the charge on the capacitor. After closing the switch, we get an ODE (11) as follows from the circuit under load, with the Eq. (9).

$$\begin{array}{l} \text{Homogeneous equation:} \quad \rightarrow \frac{dq_h(t)}{dt} + \frac{1}{RC}dq_h(t) = 0 \rightarrow \\ \rightarrow q_h(t) = ke^{-\frac{t}{RC}}, \end{array} \tag{12}$$

$$\begin{array}{c}\text{Complete equation:} \qquad \rightarrow q(t) = q_p(t) \rightarrow \\ = \text{constant} = 0 + \frac{1}{RC} q_p = \frac{1}{R} v_0 \rightarrow q_p = Cv_0.\end{array} \tag{13}$$

$$\boxed{q(t) = ke^{-\frac{t}{RC}} + Cv_0}, \tag{14}$$

for $t = 0 \rightarrow q(t = 0) = 0 \rightarrow 0 = ke^{\frac{0}{RC}} + Cv_0 \rightarrow \boxed{k = -Cv_0,}$

$$\boxed{q(t) = -Cv_0 e^{-\frac{t}{RC}} + Cv_0} \rightarrow \tag{15}$$

$$\rightarrow \boxed{q_c(t) = Cv_0 \left(1 - e^{-\frac{t}{RC}}\right),} \tag{16}$$

once $q_c(t) = Cv_c(t)$, we obtain $\qquad \boxed{v_c(t) = v_0 \left(1 - e^{-\frac{t}{RC}}\right)}$, (17)

given that $i(t) = \frac{dq_c(t)}{dt}$, we get $\qquad i(t) = 0 - Cv_0 e^{-\frac{t}{RC}} \left(-\frac{1}{RC}\right)$,

$$i(t) = \frac{v_0}{R} e^{-t/RC} = I_0 e^{-t/RC}, \tag{18}$$

To solve the ODE (11) we need to get the general solution of the homogeneous EDO associate to Eq. (11) and a particular solution of the complete Eq. (11).

The general solution to the given ordinary differential equation is the sum of solutions (12) and (13), resulting in Eq. (16) which relates equalities (17) and (18), confirming Eqs. (5) and (6). The solutions given by formulae (17) and (18) are displayed in Fig. 6.

In the first phase of this work, three capacitors were used, each with $2200\mu F$ and a maximum voltage of $385V$. With these capacitors, a parallel capacitor bank was assembled, that is, the total capacitance becomes the sum of the 3 capacitances, maintaining the voltage, $C_{total} = C_1 + C_2 + C_3$. The first problem that arose was the charging of the capacitors, since it is necessary to use a source of at least 385 V for the capacitors to get close to that value. This equipment is not commonly found in the laboratory and is perhaps expensive. Therefore, a high-voltage capacitor charging circuit was acquired, whose function is to obtain an adjustable output voltage from this circuit, between the 250 V and 420 V. The complete charging circuit is represented in the diagram on left in Fig. 7.

This charging circuit is powered by a **DC** voltage of $3.2V$ to $6V$. Before testing the right loop, calculations were made to ensure that the energy dissipated in resistor R_1 did not exceed its limit, and consequently melt or damage the rest of the components. It is known that the dissipated power is given by $P = VI$ (W ou Js_{-1}), and rewriting the current as $I = V/R$, we obtain $P = V^2/R$.

Generally, in the case of laboratory resistors, these dissipate up to $1W$. We can then express the value of R_1 so that it does not exceed its dissipation limit, thus obtaining the value of $R_1 = 160k\Omega$.

Charging Time. With the charging equations, the voltage across the capacitor at each instant is easily known, as well as the maximum current that will flow through the circuit when switch 2 is closed. In Equation (4.2), the maximum current value is reached at time $t = 0$, since the solution is simply equal to the peak current, 0, (see Fig. 4.2). It is also known that it is given by $I_0 = \frac{V_0}{R}$, so if $R_1 = 160k\Omega$ and $V_0 = 420V$, the maximum current value that flows through

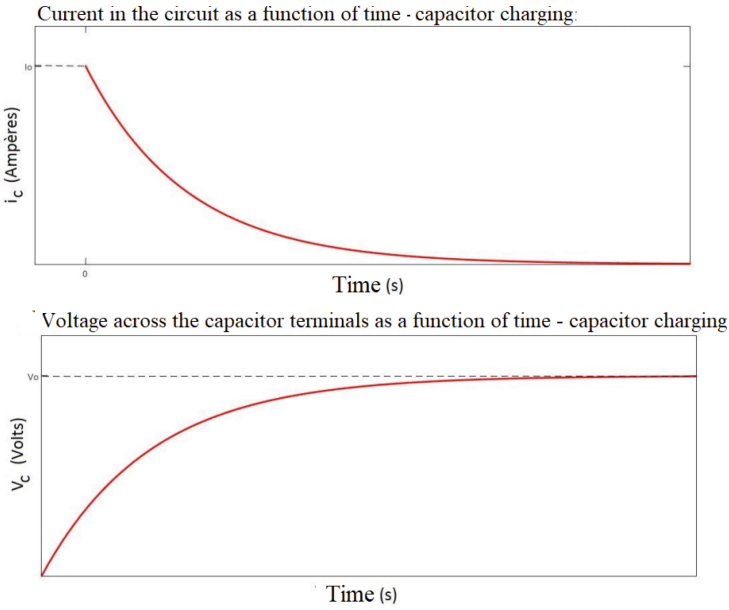


Fig. 6. Current and voltage curves as a function of time, respectively.

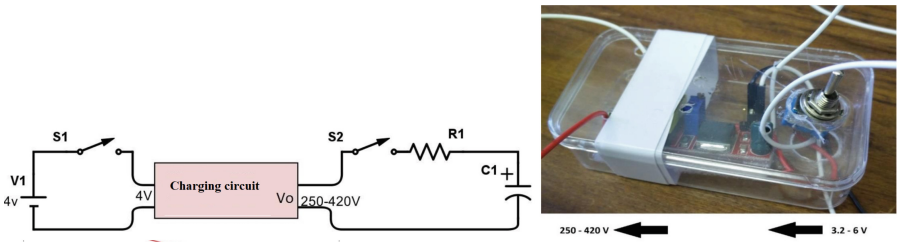


Fig. 7. Left - Charging circuit scheme. Right - Charging circuit purchased. Subsequently, this circuit was adapted and placed inside the plastic box, and a switch was added for safety purposes.

the right loop is $2.63mA$. To get an idea of how long it takes a capacitor to charge, Equation (4.1) was used. For example, to charge a capacitor at $200V$ using a $V = 400V$ and $R = 160k\Omega$, simply solve Equation (5) in order to time. Therefore, a capacitor takes about 4 min and 20 s to charge to $200V$. Of course, there is still room to reduce the charging time of the capacitors. To do this, simply lower the resistance value, however, without exceeding its power dissipation limit. A rheostat with $R = 360\Omega$ was used, which can handle currents up to $10A$. It was replaced by resistor R_1 of $160k\Omega$, thus resulting in a charging of the 3 capacitors to $200V$ in 1.54 s, theoretically. This significant improvement made it possible to carry out several laboratory tests without the constraint of the charging time.

Despite the agreement of the previous image (Fig. 8 top) with the theoretical charging curves, the actual charging times were different from the theoretical ones, namely much longer. As can be seen in Fig. 8, the charging time of 3 capacitors at $200V$ was approximately 120 seconds.

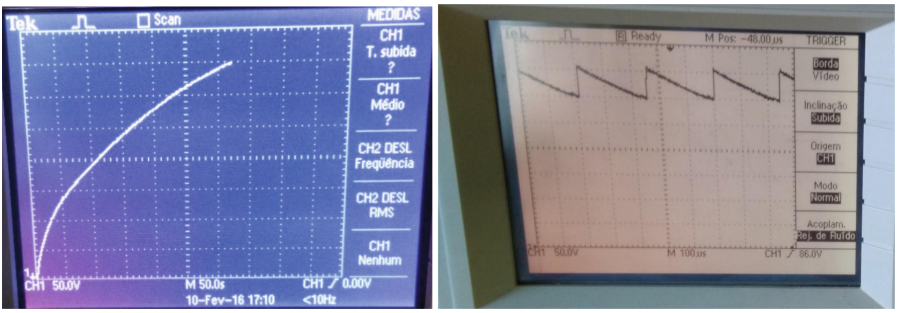


Fig. 8. Top - Image taken from the oscilloscope during capacitor charging. This image is consistent with the expected theoretical curve in Eq. (5) and Fig. 4. Bottom - Signal at the output of the capacitor charging circuit measured on an oscilloscope.

After investigating the charging circuit, it was noted that the connections between the terminals of the charging circuit and the capacitors were not the best; some contacts had dirt and even some rust, as they were quite old. These factors contributed to an increase in the overall resistance of the right-hand loop (see Fig. 8, top), and consequently lower current values and longer charging time. After cleaning the contacts, the charging time improved slightly, however, far from the theoretical values ($1.54s$). Finally, it was decided to connect the terminals of the charging circuit to an oscilloscope to see what type of signal was entering the capacitors, and the image at bottom of Fig. 8 was observed. As can be observed, we are not dealing with a clean DC signal, as one would expect, but rather a signal that oscillates up to $50Vof$ voltage. This phenomenon is called *ripple* voltage.

This *ripple* is derived from the incomplete suppression of an alternating current, created in the charging circuit, typically associated with chargers that use current rectifiers, as is the case here. The operating principle of a charger that uses current rectifiers can be described by the diagram displayed in Fig. 9.

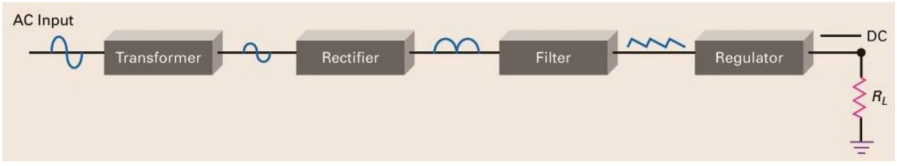


Fig. 9. Diagram of the operation of a charger based on rectifiers. Source [14].

One of the major steps that the railgun project will need is effectively a power supply specifically designed for charging capacitors at $400V$, with the capacity to charge a bank of capacitors consisting of 10 capacitors of $6600\mu F$.

6.2 Discharge Phase

The most important phase of a railgun is the discharge, as this is where all the weapon's kinematics take place. The discharge circuit can be represented by the scheme in Fig. 10.

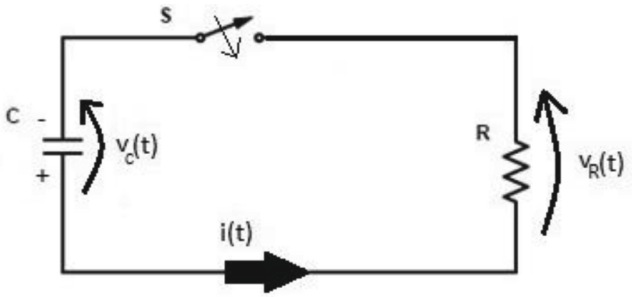


Fig. 10. Railgun discharge circuit.

Similar to the charging phase, the solutions can be obtained from the first-order differential equation of a discharging RC circuit (Eq. 21). This EDO is deduced applying the Kirchhoff, obtaining the relation defined by Equation (19). Ssolving Equation (21), aftr some calculations that follows we obtain the two solutions of interest: Equations (25) and (26).

$$-v_C(t)+v_R(t)=0 \tag{19}$$

with $v_c(t)=\frac{1}{C}q_c(t)$, $v_R(t)=Ri(t)$ and $i(t)=-\frac{dq_c(t)}{dt}$, so :

$$-\frac{q_c(t)}{C}+Ri(t)=0\rightarrow-\frac{q_c(t)}{C}+R\frac{-dq_c(t)}{dt}=0\rightarrow \tag{20}$$

$$\rightarrow\frac{q_c(t)}{C}+R\frac{dq_c(t)}{dt}=0, \tag{21}$$

$$\frac{dq_c(t)}{dt}+\frac{1}{RC}q_c(t)=0\rightarrow\boxed{q_c(t)=ke^{-\frac{t}{RC}}}, \tag{22}$$

in $t=0$ we have $v_c(t=0)=v_{c_0}$, also $q_c(t=0)=Cv_{c_0}=ke^0=k\rightarrow\boxed{k=Cv_{c_0}}$
so:

$$q_c(t\geq0)=Cv_{c_0}e^{-\frac{t}{RC}}\rightarrow \tag{23}$$

$$\rightarrow v_c(t\geq0)=\frac{1}{C}q_c(t\geq0)\rightarrow \tag{24}$$

$$\rightarrow v_c(t\geq0)=v_{c_0}e^{-\frac{t}{RC}}. \tag{25}$$

By definition, $i(t)=-\frac{dq_c(t)}{dt}=-\frac{d}{dt}\left(Cv_{c_0}e^{-\frac{t}{RC}}\right)=Cv_{c_0}e^{-\frac{t}{RC}}\frac{1}{RC}$
Simplifying, we obtain:

$$\boxed{i(t)=\frac{v_{c_0}}{R}e^{-\frac{t}{RC}},} \tag{26}$$

$$v_c(t)=v_{c_0}e^{-t/RC}, \tag{27}$$

$$I(t)=I_oe^{-t/RC}, \tag{28}$$

Summarizing, the solutions that describe the discharge curves of the capacitors are given by formulas (27) and (28), with $I_0=V_0/R$whose graphical representation is made in Fig. 11.

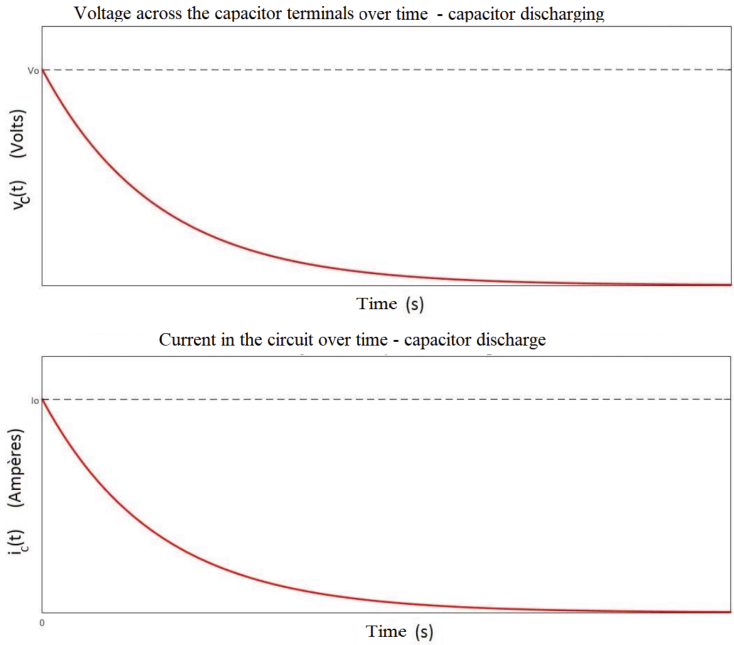


Fig. 11. Discharge curves of capacitors, above the voltage curve as a function of time Equation, below the current curve as a function of time Equation.

This is the typical behavior of a capacitor; however, for our system, we have to complete the discharge equations to make them as similar as possible to reality. To do this, it is necessary to analyze the behavior of the electrical resistance of the system.

7 Experimental Part

This is the typical behavior of a capacitor; however, for our system, we have to complete the discharge equations to make them as similar as possible to reality. To do this, it is necessary to analyze the behavior of the electrical resistance of the system. At first glance, the basic principle of operation of a railgun may seem simple in the light of the immediate application of the Lorenz force in a conductor driven by current. However, everything becomes more complicated in the case of a rapid change of the parameters involved (transitional regime), which requires a more thorough analysis of current behavior, electric and magnetic fields, and all materials involved in this system. This part of study can be found in [15], where the theoretical part is developed. In this paper an analysis in the time domain, of the spatial distribution of the energy flow was performed shortly. Now, the work is complemented by the experimental part. This project also involved the construction of two railguns, initially a smaller size one to gain familiarity with the system, which is described in [16]. Lastly, a laboratory proportion one in which was performed and analyzed a number of tests with different materials and projectile sizes that is presented here.

7.1 Railgun with Copper Rails

When building a railgun, the first concern must be energy storage, optimizing it and designing the rail architecture (length of the rails) as a function of the capacitor bank discharge time.

Ideally, the rails should have a length such that the instant the projectile leaves the rails coincides with the instant the capacitors are discharged.

We describe the construction and experimentation of one of the railguns used in this study. Due to a lack of experience in handling high currents, a smaller and simpler model of a railgun was first built in order to gain familiarity with this type of system: the railgun with aluminum rails.

The first railgun, built for this study, consisted of: a laboratory DC power supply, the capacitor charging circuit (discussed in the fourth chapter), three $2200\mu F/385V$ capacitors, two aluminum rails, and spherical steel projectiles of three different diameters. The detailed description of the results associated to the first railgun can be found in [16].

In the analysis of the slow-motion images of the shots, although the demonstration of the Lorentz force is quite visible, the efficiency of the first railgun was very low. The existence of sparks flying after the discharge and the formation of arcs were the causes of the low efficiency, as the waste the energy that should be used to accelerate the projectile. That project worked with spheres as projectiles, it was suspected that such geometric shape was not ideal for obtaining good electrical contact. The contact sections must be compatible, that is, if the rails are rectangular in section, the projectile must also have flat sections. To lower the electrical resistance and, consequently, increase the current in the system, copper rails were used in the actual project since they have a higher electrical conductivity. The projectile pre-acceleration system was adapted because we no longer have spheres as projectiles, this ramp will cease to function. The second railgun project are made of copper and had a rectangular cross-section, thus, we improved not only the contact between rails and projectile (both have flat cross-sections) but also the electrical resistance of the projectile due to the greater conductivity of copper compared to steel With these changes, it is expected that efficiency will increase, making the Lorentz force effect more visible, and also that the railgun will be capable of multiple shots without the need for rail repair.

For the construction of the second railgun, copper rails were used. The choice of this material is due to its high electrical conductivity, $5.91 \times 10^{-9}(\Omega m)^{-1}$, almost twice as good as aluminum, $3.63 \times 10^{-9}(\Omega m)^{-1}$, used previously. It was decided to use rails with $50cm$ each, as this length allows a complete discharge of the $106600\mu F$ capacitors in parallel, according to the study done in the fourth chapter. It was agreed that the spacing between rails would be $4mm$ (which will also be the width of the projectiles), however, adjustable if projectiles with different dimensions are to be used. Projectiles with a width of $4mm$ and different lengths were used in order to perform various tests and determine which one yields the best results (see Fig. 12).

To reduce the danger of using the railgun, an insulating section was created at the beginning of the rails, in order to allow the projectile to be placed inside the rails safely.

Acrylic was used as support for the structure instead of wood, which was used previously. A sheet of acrylic was placed on top and another underneath, with the rails in the middle of the sheets. This prevents pieces of molten metal from flying towards the operator, as happened in the previous project. Both the acrylic and the rails are fixed by 8 screws and 8 nuts located asymmetrically, preventing the rails from separating during firing due to the Lorentz force that also manifests itself between these.

7.2 Details of the Shooting

Four sets of tests were performed with the following initial conditions:

- With 1 capacitor bank of 3 capacitors at 350 V of 2200 μF, without lubricant;
- With 1 capacitor bank of 3 capacitors at 200 V of 2200 μF, with graphite on the rails;
- With 1 capacitor bank of 4 capacitors at 200 V of 6600 μF, without lubricant;
- With 1 capacitor bank of 2 capacitors at 380 V of 6600 μF, with a silver-based grease.

In this second railgun, another form of projectile pre-acceleration was tested.

A small slingshot was constructed, consisting of a plastic tube, a balloon, a thin rubber band, and a thin bamboo stick. In the first set of tests, 3 shots were fired, one successfully, hitting a styrofoam plate that was 20cm from the end of the rails, but two misfires occurred in the last shots. After the misfire, it was found that there was already some wear on the rails. In order to reduce wear on the rails, in the second set of tests, graphite was used as a conductive lubricant on the rails and projectile, similar to that used by Pavlov [5]. To apply the surface layer of graphite, a conventional pencil was used, scraping it against the surface of the rails. In the 3 shots performed, two were successfully recorded (see Fig. 12), similar to the shots in the first set of tests.



Fig. 12. Left - Comparison of the length of various copper projectiles. From top to bottom, length $L = 2.0cm$, $L = 1.8cm$ and $L = 0.6cm$. Right -Frame during the first shot of the second round of testing.

The recorded misfire occurred because the projectile welded to the rails in the first moments of discharge, resulting in considerable excavation in the rails. The major difference observed from the first set of shots to the second was the reduction in rail wear, which is an improvement in the longevity of the material. For the third round of tests, we already had the new 6600 $\mu F/400V$ capacitors available. Four capacitors were mounted in parallel, as shown in the figure. With this setup, the Lorentz force will be of the same intensity as in the second set of tests (see Eq. 3), since we are using the same voltage (and consequently the same initial current[2], on the order of tens of thousands *Amperes*), however, this force will be exerted for a longer time, namely, four times longer due to the increased capacitance of the new capacitor bank (Fig. 13).

Three shots were fired, none of them successful, resulting in a more pronounced excavation than in previous cases. It should be noted that in this case we had four times more energy involved in the discharge, which caused the projectile to heat up even more and weld itself to the rails, damaging both in such a way that it was imperative to carry out repairs to the rail, eliminating the irregularities caused by the shots.

Analyzing the firing failures of the previous tests, it was concluded that the problem was possibly at the level of friction that the projectile encountered while sliding between the rails. We tried to find a lubricant (similar to graphite) that would reduce friction but not increase electrical resistance. We were advised by a specialist to use a silver39-based grease, which a prior met the

[2] Since during the discharge of capacitors, we have $I_0 = \frac{V_0}{R}$ and $I(t) = I_0 e^{-\frac{t}{RC}}$ ..



Fig. 13. Capacitor assembly used in the third and fourth rounds of railgun testing.

aforementioned requirements. After applying the lubricant to the rails , it was noted that the mass was somewhat dense, which could increase friction instead of decreasing it.

For the final round of tests, in order to give the projectile an initial velocity and prevent it from sticking to the rails, another method was used to provide initial propulsion to the projectile. Compressed air similar to that used for cleaning hardware in electronics was employed. Instead of 4 capacitors in parallel, it was decided to use 2 capacitors in parallel, which allowed for a reduction in the charging time of the capacitors and the use of a higher voltage, 380 V. Three shots were made with the applied lubricant, and similar to the tests in the previous round, they resulted in misfires. The projectile ended up sticking to the rails at the beginning of the discharge, and it was observed that the damage done to the rails was even more severe than in previous cases (see Fig. 14), both in terms of rail deformation and the excavation caused by the projectile, ultimately rendering the rails unusable for future tests (Table 2).

Table 2. Data from the results obtained after the four sets of railgun shots.

	Voltage	Capacitance	successfull shoots	unsuccessfull shoots
No lubricant	350 V	6600 µF	1	2
Graphite	250 V	6600 µF	2	1
No lubricant	200 V	26 400 µF	0	3
lubricating grease (silver)	380 V	12 000 µF	0	3

7.3 System Damage

Two types of damage were recorded in the system, mentioned in the first chapter: the excavation in the rails and the effects of current arcs on the projectiles. The excavation in the rails became increasingly greater as the amount of energy in the capacitors increased, being more visible in the last round of firing. The damage caused by the excavation led to the rails becoming unusable after 12 shots, as can be seen in the Fig. 14



Fig. 14. Damage to the rails after the last round of shots. Damage caused by the excavation effect of the projectile on the rails. Damage caused by arcing currents in the initial zone (right) of the projectile. Note the similarity of the wear pattern to that found on other railguns.

The lubricant that yielded the best results was graphite, both in terms of material protection and weapon effectiveness (fewer misfires). Initially, it was thought that grease would yield better results, however, due to its high density and viscosity, it ended up further increasing the coefficient of friction between the projectile and the rails, leading to misfires. Despite the material being of high quality, it was not the most appropriate considering the reduced dimensions of the railgun. It should be noted that graphite is much less expensive (common in traditional graphite pencils) and easier to apply compared to grease, which requires using the precautions recommended by the manufacturers for its application.

After several shots, it was concluded that the larger projectile, obtained the best results. The smaller, lighter projectiles ended up generating the recorded misfires. The larger projectile, perhaps because it has greater inertia, resists the initial friction between the rails and it projects itself better. Due to the wear of the projectile, it is not advisable to use it for more than three shots, since the deposits accumulated on the projectile, in addition to adding irregularities, have low conductivity. When the irregularities in the projectile are removed, strictly speaking, we are reducing the dimensions of the projectile. These small changes are enough to create a slight misalignment between the rails and the projectile that could increase the probability of excavation and current arcs.

At the capacitor level, numerous discharge tests were performed (more than 30 discharges, in addition to the recorded discharges) for a rheostat, which functioned as a load, with no signs of wear on the capacitors or decrease in their performance.

8 Discussion and Final Remarks

Major Constraints

There were two major constraints that limited the analysis to the obtained results. The first was the absence of equipment to measure the projectile's velocity at the muzzle of the rails. With this measurement, it would be possible to have a general idea of the weapon's performance, comparing velocities with different initial conditions, and also to compare the practical results with the theoretically predicted results. The second limitation was the lack of a clamp or hook anemeter, which would allow quantifying the current present during firing; if we had this measurement, it would be possible to predict the theoretical velocity of the projectile, which, when compared with the actual velocity, would give the efficiency index η of the weapon.

Rails Size

An important conclusion drawn was regarding the size of the rails, where it was verified that the discharge of the capacitors was carried out only with the projectile in the first third of the tracks, that is, the rest of the track length contributed only with friction. Although the

total capacitance of an assembly of $106600\mu F$. capacitors in parallel, which would extend the discharge of the bank, was not used, it is predicted that half the current length, $L_{rail} = 25cm$, would be sufficient to complete the discharge of the capacitor bank proposed.

Capacitor Bank and Power Supply

In the tests performed, only four of ten $6600\mu F$ capacitors were used, as the charging circuit could not handle such high currents. Furthermore, when testing the charging of six capacitors in parallel, some of the circuit components melted and smoke was released. This will be a problem that will inevitably have to be addressed in future designs of this railgun, as we aim to charge the capacitor bank proposed above (ten capacitors). It should also be noted that the development of a power supply designed for the fast charging (less than 60 s) of a capacitor bank at a voltage of 1200 V with a capacitance of $26400\mu F$.

Regarding the capacitor bank configuration, with the proposed change to the rails (reduction to 25 cm in length), it is proposed, similar to the discussion in the fourth chapter, that tests be carried out with a configuration of a bank consisting of three secondary banks mounted in series, in which each secondary bank has four capacitors connected in parallel, resulting in a current of 19 kJ with a discharge time of $9.6\mu s$, similar to what was done in the Powerlabs railgun [6]. With this configuration, an optimization between the relationship between "discharge time" and "available energy" is expected.

Projectile Dimensions

Although several projectile dimensions were tested for the railgun, it was not conclusive what the optimal projectile length would be; however, it was confirmed that the largest dimension, $L = 2.0cm$, obtained better results compared to the smaller ones. It is encouraged that other materials be tested for the constitution of the projectile, particularly aluminum, which presents itself as a strong candidate due to its electrical conductivity, low density, and results already obtained in other projects [1,6].

The projectile's architecture itself needs reforms, as the electrical contact seems to be made only in the tail region of the projectile, not taking advantage of the entire lateral surface of the projectile to conduct current, as can be seen in Fig. 15. General Atomics, the most advanced company in this technology, proposed a geometry for its projectile, encapsulating it in a sabot, in which the contact zone with the rails has the shape of a "U" (see Fig. 15 [17]. Other geometry has already been adopted by other railgun designs.

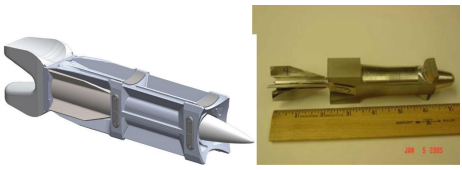


Fig. 15. On the left, the geometry adopted by General Atomics, and on the right, by the EM Gun Project.

Acknowledgements. The authors ware supported by Portuguese funds through the *Center of Naval Research* (CINAV), Naval Academy, Portuguese Navy, Portugal.

References

1. Schoeder, M.K.: An Investigation of the Static Force Balance of a Model Railgun. Monography, Naval Postgraduate School, Monterey, USA (2007)
2. Graneau, P.N.G.: Newtonian Electrodynamics. World Scientific, Singapore (1995)
3. ourselfGadjets, D.I.: Homemade railgun (2013)
4. Persad, C.: Characterization of hypervelocity gouge craters in rail conductors. IEEE Trans. Magn. **33**(1), 197–202 (1997)
5. Pavlov, A.: Improvement of Electromagnetic Railgun Barrel Performance and Lifetime by Method of Interfaces and Augmented Projectiles. Monography, Faculty of California Polytechnic State University, California, USA (2013)
6. PowerLabs: Powerlabs rail gun 2.0 research! (2010). http://www.powerlabs.org/railgun.html. Accessed 20 Apr 2026
7. Maniglia, J.: Design, fabrication, and testing of an eletromagnetic rail gun for the repetead testing and simulation of orbital debris impacts (2013). http://digitalcommons.calpoly.edu/aerosp/. Accessed 20 Apr 2026
8. Bauer, D.P.: Achieving high efficiency with conventional railgun launchers. IEEE Trans. Magn. **31**, 263–266 (1995)
9. Bauer, D.P.: A novel railgun launcher design. IEEE Trans. Magn. **31**, 267–272 (1995). https://api.semanticscholar.org/CorpusID:111118679
10. Herald, T.S.M.: Us navy's new destroyer zumwal will test fire futuristic eletromagnetic railgun (2015). http://www.smh.com.au/technology/sctech/us-navys-new-destroyer-zumwalt-will-test-fire-the-futuristicelectromagnetic-rail-gun-20151209-gljao4.html. Accessed 20 Apr 2026
11. Bindu., S., et al.: Modelling of a spark gap switch. In: 2012 IEEE 10th International Conference on the Properties and Applications of Dielectric Materials, pp. 1–4. IEEE (2012)
12. Maniglia, J.: Design, Fabrication, and Testing of an EMR Gun based Orbital Debris Impact Testing Platform. Faculty of California Polytechnic State University, California, USA, Monography (2013)
13. Atomics, G.: General atomics team powers navy rail gun to new world record (2013)
14. Wendling, M.: Retificadores. Universidade Estadual Paulista, São Paulo, Brasil (2025). http://www2.feg.unesp.br/Home/PaginasPessoais/ProfMarceloWendling/3---retificadores.pdf. Accessed 20 Apr 2026
15. Sampaio, P., Gatta, M., Teodoro, M.F.: Studying of the operation of a railgun system: a theoretical approach. Mathematics (2026)
16. Sampaio, P., Gatta, M., Teodoro, M.F.: Investigating of the operation of a railgun system: experimental approach. In: Machado, J.E.A. (ed.) Lecture Notes in Mechanical Enginneering. Springer (To Appear)
17. Atomics, G.: Projectile development (2016). http://www.ga.com/railgunprojectile-development. Accessed 20 Apr 2026

Determinants of Housing Prices in Portugal: A Municipal-Level Analysis

Maria J. Ferreira[1], Aldina Correia[1(✉)], Vanda Lima[1], and Diogo Ribeiro[2,3]

[1] CIICESI, ESTG, Instituto Politécnico do Porto, Rua do Curral, Casa do Curral, Margaride, 4610-156 Felgueiras, Portugal
aic@estg.ipp.pt

[2] Department of Data Science and Research, MySense, London, UK

[3] ESMAD, Instituto Politécnico do Porto, Rua D. Sancho I, n.o 981, 4480-876 Vila do Conde, Portugal

Abstract. This study analyzes the determinants of housing transaction values at the municipal level in Portugal, providing empirical evidence to support decision-making by households, investors, and policymakers. Using cross-sectional data for all 308 Portuguese municipalities in 2017, obtained from the PORDATA database, the analysis estimates a multiple linear regression model with the logarithm of the average value of housing transactions as the dependent variable. Explanatory variables capture geographic location, demographic characteristics, economic capacity, labor market conditions, environmental quality, and housing market structure.

The results indicate that regional location, purchasing power per capita, average family size, homeownership rates, and housing availability are the most significant determinants of municipal housing prices. In contrast, population density and environmental quality do not exhibit statistically significant effects once economic and structural factors are controlled for. Robustness checks using Generalized Least Squares confirm the stability of the main findings.

By focusing on the municipal scale and jointly considering demand-side and supply-side factors, this study contributes to the literature on housing price formation in Portugal and complements existing national and regional analyses. The findings highlight the importance of accounting for regional heterogeneity and housing supply conditions when designing policies aimed at improving housing affordability and market efficiency.

Keywords: Housing Price · Portugal · Multiple Linear regression · Decision making · Housing policy · Data analysis

1 Introduction

Housing markets play a central role in economic stability, household welfare, and social cohesion. The global financial crisis of 2007–2009 highlighted the strong

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 87–98, 2027.
https://doi.org/10.1007/978-3-032-30521-3_6

interconnections between housing prices, credit conditions, and macroeconomic performance, reinforcing the relevance of housing markets for both economic policy and financial stability ([7,22]). Since then, housing affordability and price volatility have become persistent concerns in many countries, particularly in urban and coastal regions where demand pressures coexist with constrained supply.

A substantial body of literature has examined the determinants of housing prices, identifying income, credit conditions, demographic factors, and location-specific attributes as key drivers of housing market dynamics. Cross-country evidence shows that housing prices are closely linked to macroeconomic and financial variables, such as income growth, interest rates, and credit availability, while institutional and structural differences explain observed heterogeneity across countries [22]. At the regional and urban level, hedonic pricing models consistently highlight the importance of location, accessibility, and structural dwelling characteristics in explaining spatial price variation ([1], [2]) and [16].

Recent studies further emphasize that supply-side constraints and institutional arrangements play a decisive role in shaping long-term housing price trajectories. Analyses of highly constrained markets, such as Hong Kong, show that persistent limitations on land supply and housing availability can sustain high price levels despite repeated policy interventions [6]. Similar concerns have emerged across European housing markets, where limited supply responsiveness has amplified price pressures and affordability challenges in recent years [10]. These findings suggest that housing price formation reflects a complex interaction between demand-side forces and structural characteristics of housing markets.

In the Portuguese context, existing research has primarily focused on national or regional dynamics over longer time horizons, highlighting the influence of macroeconomic conditions, credit cycles, and demographic trends on housing prices [17,23]. Other studies employing hedonic approaches at a regional or metropolitan scale confirm the relevance of location and local public dynamics in shaping housing values ([13,19]). However, empirical evidence at the municipal level for the entire Portuguese territory remains relatively scarce, particularly studies that jointly consider demographic, economic, social, and housing supply factors within a unified econometric framework.

This paper seeks to address this gap by analyzing the determinants of housing transaction values across all 308 Portuguese municipalities using cross-sectional data for 2017. By combining location indicators, demographic characteristics, economic capacity measures, and housing market structure variables, the study provides a comprehensive assessment of the factors driving municipal-level housing price variation. The empirical analysis employs multiple linear regression techniques and robustness checks to identify the variables most strongly associated with housing transaction values. The results aim to contribute to the literature by offering new evidence on housing price determinants at a fine spatial scale and by informing policymakers and market participants about the structural forces shaping housing market outcomes in Portugal.

Therefore this study examines how municipal-level characteristics influence housing values in Portugal, addressing a gap in the empirical literature on local housing markets. By quantifying the impact of demographic, economic, and housing-related factors on transaction values, the analysis provides evidence relevant for households, investors, and policymakers, supporting more informed decision-making and housing policy design.

The remainder of this paper is organized as follows. Section 2 reviews the main theoretical and empirical contributions in the literature on housing price determinants, with particular emphasis on location, demographic, economic, and housing supply factors. Section 3 presents the econometric framework, describing the Ordinary Least Squares and Generalized Least Squares methodologies adopted in the analysis. Section 4 introduces the data, outlines the model specification, and discusses the empirical results at the municipal level. Finally, Sect. 5 concludes the paper, summarizing the main findings and outlining policy implications and directions for future research.

2 Factors Influencing the Value of Housing

The value of housing is shaped by a complex interaction of geographic, demographic, economic, and social factors. A substantial body of literature has investigated these determinants, commonly employing hedonic pricing models and regression-based approaches to explain spatial and temporal variations in housing prices. However, empirical evidence at the municipal level remains relatively limited, particularly in the Portuguese context, which motivates the present study.

Location is consistently identified as one of the most influential determinants of housing values. Accessibility, proximity to employment centers, public services, and environmental amenities contribute significantly to spatial price differentials [1]. Studies have shown that quality-of-life attributes, including access to green spaces, coastal areas, and urban infrastructure, are capitalized into housing prices, reflecting households' willingness to pay for favorable living conditions [5]. Empirical analyses conducted in metropolitan regions such as Istanbul, Ankara, and several Spanish provinces further confirm the dominant role of location in explaining housing price variation [2,18]. These findings justify the inclusion of territorial classifications, such as NUTS levels, as key explanatory variables.

Demographic pressure, commonly proxied by population density, has also been associated with housing market dynamics. Zhou et al. [24] documented a positive relationship between population density and housing prices across Chinese cities, suggesting that denser urban environments tend to exhibit higher demand and increased productivity. Similarly, Rappaport [20] found that higher density areas in the United States experience rising land and housing prices, reflecting the capitalization of urban amenities into property values. These results support the hypothesis that population concentration influences housing demand and price formation.

Economic conditions represent another central dimension in housing valuation. Income and purchasing power are repeatedly shown to be strong predictors

of housing prices, as they directly affect households' ability to acquire property. McGreal and Taltavull de La Paz [18] emphasized the relevance of income and accessibility factors in determining housing prices across European regions. In addition, Fuller et al. [11] highlighted the long-term implications of housing price dynamics for wealth accumulation and intergenerational equity. Accordingly, per capita purchasing power is included in the model as a fundamental economic determinant.

Household composition and labor market conditions further contribute to housing market behavior. Family size influences the demand for housing types and dwelling sizes, with larger households typically requiring more spacious accommodations. Empirical evidence from urban housing markets, such as the Chicago metropolitan area, confirms the importance of family size in explaining housing prices [3]. Unemployment rates, on the other hand, reflect local economic vulnerability and income uncertainty. Johnes and Hyclak [14] demonstrated that labor market conditions, including unemployment, are closely linked to housing prices through their impact on purchasing capacity and residential stability.

Social and demographic characteristics, including population aging and educational attainment, also play a role in shaping housing demand. Aging populations may alter housing preferences and slow market dynamics, while higher education levels are often associated with higher incomes, greater residential mobility, and increased housing demand [4,8]. These relationships motivate the inclusion of aging and education indicators as explanatory variables.

Finally, housing market structure itself influences price behavior. Homeownership rates may signal market maturity or saturation, while housing availability affects supply-side pressures. Studies on housing supply constraints indicate that limited availability can significantly increase prices, particularly in regulated or densely populated regions [15]. Consequently, indicators capturing owner-occupied housing and dwelling availability are incorporated to account for structural market effects.

Overall, the literature supports a multifactorial approach to housing price modeling, where location, demographic pressure, economic capacity, household structure, and housing supply jointly influence transaction values. Building on these insights, the present study empirically examines whether these determinants exhibit similar effects at the municipal level in Portugal.

3 Model Theoretical Framework

This section presents the econometric framework adopted to analyze the determinants of housing prices in Portuguese municipalities. The methodological approach is based on linear regression techniques, starting with Ordinary Least Squares (OLS) estimation and extending to Generalized Least Squares (GLS) to address potential violations of classical assumptions.

3.1 Ordinary Least Squares Estimation

Ordinary Least Squares (OLS) is a standard estimation technique for multiple linear regression models and serves as a natural starting point for empirical housing market analysis. The method estimates the parameters of a linear relationship between a dependent variable and a set of explanatory variables by minimizing the sum of squared residuals.

Model Specification. The general multiple linear regression model is defined as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations of the dependent variable, $\mathbf{X}$ is an $n \times k$ matrix of explanatory variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of error terms.

In this study, the dependent variable corresponds to the logarithm of the average value of housing transactions (VMPT), while the explanatory variables represent location, demographic, economic, and housing market characteristics at the municipal level.

Estimation and Properties. The OLS estimator of the parameter vector $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Under the Gauss–Markov assumptions, this estimator is the Best Linear Unbiased Estimator (BLUE). These assumptions include linearity of the model, absence of perfect multicollinearity among regressors, zero conditional mean of the error term, homoscedasticity, and absence of serial correlation.

Model adequacy is commonly assessed using the coefficient of determination R^2 and its adjusted version, which accounts for the number of explanatory variables. In addition, the overall statistical significance of the model is evaluated using the F-test.

Assumptions Verification. The validity of OLS inference relies on the verification of its underlying assumptions. Residual diagnostics are therefore conducted to assess normality, homoscedasticity, and serial independence. Multicollinearity among regressors is examined using correlation analysis and variance inflation factors, allowing the identification and exclusion of redundant explanatory variables when necessary.

Interpretation of Coefficients. The estimated regression coefficients represent the expected average change in the dependent variable associated with a one-unit change in the corresponding explanatory variable, holding all other variables constant. Given the logarithmic transformation of the dependent variable, the coefficients can be interpreted as semi-elasticities, facilitating economic interpretation in terms of percentage changes in housing values.

3.2 Generalized Least Squares Estimation

Although OLS provides consistent and unbiased estimates under ideal conditions, empirical housing data often exhibit heteroscedasticity and autocorrelation, particularly in cross-sectional datasets with spatial or regional structure. When such violations are detected, OLS estimators remain unbiased but lose efficiency, and standard errors may be inconsistent.

To address these issues, the analysis is extended to a Generalized Least Squares (GLS) framework. GLS relaxes the assumption of spherical error terms by allowing for a non-diagonal variance–covariance matrix of the disturbances. The GLS estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y},$$

where $\boldsymbol{\Omega}$ denotes the variance–covariance matrix of the error term.

By explicitly modeling the structure of the error variance, GLS yields more efficient parameter estimates and reliable inference when heteroscedasticity or autocorrelation is present. In this study, GLS is employed as a robustness check to confirm the consistency of the estimated effects obtained under OLS.

4 Data and Results

4.1 Data Description

This study is based on cross-sectional data for the year 2017, covering all 308 Portuguese municipalities, including mainland Portugal and the Autonomous Regions of the Azores and Madeira. Demographic indicators reported by the National Statistics Institute (INE) highlight a declining population trend since 2010 and a progressive aging process, reflected in an increase in the median age from 42.7 to 44.2 years between 2012 and 2017.

The empirical analysis uses data obtained from the PORDATA database, which provides harmonized municipal-level information on demographic, economic, and housing-related variables. The spatial distribution of municipalities across NUTS I and NUTS II regions is presented in Table 1. Location variables were coded as ordinal indicators to capture regional heterogeneity within the regression framework.

Housing value is measured by the Average Value of Property Transactions (VMPT). Explanatory variables include population density, purchasing power per capita, average family size, unemployment rate, environmental quality, aging index, education level, marital status, homeownership rate, and housing availability. With the exception of the location variables, all regressors are continuous. A detailed description of the variables and their measurement units is provided in Table 2.

Table 1. Distribution of Municipalities by NUTS I and NUTS II

NUTS Level	Classification	Frequency	Percentage	NUTS Level	Classification	Frequency	Percentage
I	Continente	278	90.3	II	Alentejo	58	18.8
				II	Algarve	16	5.2
				II	Lisboa	18	5.8
				II	Centro	100	32.5
				II	Norte	86	27.9
I	Madeira	11	3.6	II	Madeira	11	3.6
I	Açores	19	6.2	II	Açores	19	6.2

Table 2. Determinants of Housing Market and Their Measurements

Determinant	Description	Unit of Measurement
Housing Value/Price	Average Value of Transacted Buildings	VMPT (€)
Location	NUTS I, II, and III	NA
Population Density	Population Density	DP (People/Km^2)
Purchasing Power	Per Capita Purchasing Power	PCPC (€/person)
Family Size	Average Size of Families	DMF (People)
Unemployment	Percentage of Unemployed Registered at IEFP	DIEFP (%)
Environmental Quality	Environmental Quality Index	IQA (€/inhabitant)
Aging	Aging Index	IE (elderly per 100 young people)
Education Level	Percentage of People with Higher Education	PES (%)
Marital Status	Percentage of Married Individuals	PCAS (%)
Home Ownership	Percentage of Owner-Occupied Dwellings	PFHP (%)
Existing Dwellings	Proportion of Dwellings per Inhabitant	PFH (dwellings/inhabitant)

4.2 Model Specification and Estimation

To analyze the determinants of housing prices, a multiple linear regression model was estimated with the natural logarithm of VMPT as the dependent variable. The logarithmic transformation allows the estimated coefficients to be interpreted as semi-elasticities, facilitating economic interpretation in percentage terms.

$$\begin{aligned}\ln(\mathrm{VMPT}_i) = {} & \beta_0 + \beta_1 \mathrm{NUTSII}_i + \beta_2 \mathrm{DP}_i + \beta_3 \mathrm{PCPC}_i \\ & + \beta_4 \mathrm{DMF}_i + \beta_5 \mathrm{DIEFP}_i + \beta_6 \mathrm{IQA}_i \\ & + \beta_7 \mathrm{IE}_i + \beta_8 \mathrm{PCAS}_i + \beta_9 \mathrm{PFHP}_i \\ & + \beta_{10} \mathrm{PFH}_i + \varepsilon_i.\end{aligned} \tag{1}$$

Multicollinearity diagnostics revealed strong correlations between certain explanatory variables. In particular, education level (PES) exhibited a high correlation with purchasing power per capita (PCPC) and was therefore excluded from the final specification.

The refined model was estimated using Ordinary Least Squares (OLS). The adjusted R^2 of 0.7121 indicates that approximately 71.2% of the variation in municipal housing transaction values is explained by the included regressors. The overall model was statistically significant at conventional levels.

Diagnostic tests were conducted to assess the validity of the OLS assumptions. The Breusch–Pagan test provided no evidence of heteroscedasticity, while the Shapiro–Wilk test indicated deviations from normality. Given the large sample size, inference based on asymptotic normality remains appropriate. Evidence of autocorrelation motivated the estimation of a Generalized Least Squares (GLS) model as a robustness check.

4.3 Regression Results and Interpretation

The estimated GLS coefficients are reported in Table 3. The results indicate that location, economic capacity, household composition, and housing market structure play a significant role in explaining municipal housing prices in Portugal.

Table 3. GLS Estimation Model Coefficients

Coefficient	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	10.74110	0.894	15.149	$< 0.0001^{***}$
NUTSII	-0.12690	0.220	-9.670	$< 0.0001^{***}$
DP	0.00003	0.045	0.572	0.5674
PCPC	0.00454	0.242	4.070	$< 0.0001^{***}$
DMF	0.33442	0.217	4.616	$< 0.0001^{***}$
DIEFP	-0.01575	0.113	-1.655	$0.0989^{\cdot}$
IQA	0.35309	0.658	-1.111	0.2686
IE	-0.00021	0.503	-1.803	$0.0724^{\cdot}$
PCAS	0.41720	1.150	-4.746	$< 0.0001^{***}$
PFHP	-1.56822	0.562	-5.825	$< 0.0001^{***}$
PFH	62.04535	40.900	2.167	0.031^{*}

Notes: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$.
Residual standard error: 0.3442.
Degrees of freedom: 308 total; 297 residual.

The strong and statistically significant effect of the NUTS II location variable confirms the central role of regional heterogeneity in housing price formation. This result is consistent with earlier hedonic and regional housing studies, which emphasize the persistence of spatial price differentials driven by accessibility, economic concentration, and coastal proximity [1,18]. In the Portuguese context, similar evidence has been reported by Lourenço and Rodrigues [17], who document pronounced regional disparities in housing prices following the financial crisis, particularly between coastal/metropolitan areas and inland municipalities.

Average family size (DMF) also exhibits a positive and statistically significant effect on housing transaction values. This result supports the hypothesis that household composition influences housing demand, particularly with respect to

dwelling size and quality. Similar conclusions were reached by Archer et al. [3], who identified family size as a dominant determinant of housing prices in urban markets, reflecting increased demand for larger or more suitable housing units.

The positive and significant coefficient associated with purchasing power per capita (PCPC) aligns with a broad strand of literature identifying income and economic capacity as key drivers of housing demand. Empirical studies across European regions consistently show that higher income levels translate into higher willingness and ability to pay for housing, thereby exerting upward pressure on prices [12,18]. This finding corroborates evidence from the Portuguese housing market, where income growth and improved credit conditions have been closely associated with housing price appreciation in recent years [17].

Several variables traditionally highlighted in the literature, such as population density, unemployment rate, environmental quality, and demographic aging, were not statistically significant in the final specification. This does not necessarily imply that these factors are irrelevant, but rather suggests that, at the municipal level, their effects may be indirect or captured by stronger covariates such as purchasing power and regional location. Similar attenuation effects have been documented in regional housing models where income and supply-related variables dominate once included jointly [14,17].

Finally, the significant effects associated with homeownership rates (PFHP) and housing availability (PFH) highlight the importance of housing market structure. A higher prevalence of owner-occupied dwellings may signal market maturity or reduced transaction activity, while greater housing availability appears to alleviate supply constraints. These findings are consistent with studies emphasizing the role of supply-side factors and regulatory constraints in shaping housing price dynamics [9,15].

5 Conclusions

This paper examined municipal-level determinants of housing transaction values in Portugal using cross-sectional data for 2017. The principal empirical findings indicate that regional location (NUTS II), purchasing power per capita (PCPC), average family size (DMF), and structural housing variables (homeownership rate—PFHP—and dwellings per inhabitant—PFH) are the covariates most consistently associated with municipal VMPT. These relationships remain robust when estimated under GLS as a robustness check for heteroscedasticity and autocorrelation.

When compared with recent national and European analyses, our results align with several documented patterns. First, the dominant role of regional location and a clear coastal/metro premium in housing values is consistent with recent Portuguese analyses that emphasise strong spatial heterogeneity and supply constraints in high-demand areas [17,21]. Second, the positive and significant effect of purchasing power per capita is in line with cross-country evidence showing that income and debt capacity are primary drivers of housing demand and affordability pressures [10,12]. Third, the evidence that structural supply factors

matter (PFH and PFHP) coheres with Eurostat and Commission findings which highlight that limited supply responsiveness in many EU regions has amplified price pressures in recent years [9,10].

Some variables included in the specification—notably population density (DP) and the environmental quality index (IQA)—were not statistically significant in the final model. This suggests that, at the municipal aggregation used here, macroeconomic and structural variables (income, location, housing stock) tend to dominate price formation, while certain local attributes may operate through indirect channels or be collinear with stronger proxies. This pattern is also observed in other empirical exercises that find the relative importance of local amenities declines when stronger income and supply variables are controlled for [17].

This study contributes to the literature by providing a comprehensive municipal-level analysis for the entire Portuguese territory, combining demographic, economic, social, and housing supply variables within a unified econometric framework.

Policy implications stemming from these findings are threefold: (i) Demand-side measures only (for example, tax incentives to buy) are likely to exacerbate price pressures where supply is inelastic; targeted supply-side interventions (increasing permitted supply, accelerating permitting, subsidising affordable construction) are therefore essential [10,21]. (ii) Policies that raise local purchasing power without accompanying supply measures can have unintended distributional effects on affordability; coordinated measures balancing supply, demand and access to credit are preferable [9,12]. (iii) Regional heterogeneity matters: national policy should be combined with regionally-tailored instruments to account for the marked differences between metropolitan/coastal and inland municipalities [17].

This study is based on cross-sectional data for 2017, its main limitation; panel and spatial models would allow direct estimation of temporal dynamics, spatial spillovers and price momentum, which recent work suggests are important in Portugal's coastal and metropolitan markets [21]. Further research should also explore microdata (parish or transaction-level) and causal identification strategies (instrumental variables, difference-in-differences, or synthetic control designs) to strengthen causal claims about supply interventions and policy measures.

The concluding remark of the work is that the municipal evidence for Portugal shows that geography (NUTS II), purchasing power, and housing market structure are the main vectors of variation in transaction values. These results are broadly consistent with contemporaneous national and European analyses and underline the need for supply-focused and regionally-differentiated policy responses to restore housing affordability.

Acknowledgments. This work has been supported by national funds through FCT - Fundação para a Ciência e Tecnologia through project UIDB/04728/2020.

Disclosure of Interests. The authors declare that they have no competing interests that are relevant to the content of this article.

References

1. Adair, A.S., Berry, J.N., McGreal, W.S.: Hedonic modelling, housing submarkets and residential valuation. J. Prop. Res. **13**(1), 67–83 (1996)
2. Alkana, L., et al.: Housing market differentiation: the cases of yenimahalle and çankaya in ankara. Int. J. Strateg. Prop. Manag. **19**(1), 13–26 (2015)
3. Archer, W.R., Ling, D.C., Smith, B.C.: Ownership duration in the residential housing market: the influence of structure, tenure, household and neighborhood factors. J. Real Estate Finance Econ. **40**(1), 41–61 (2010)
4. Brasington, D.M., Hite, D., Jauregui, A.: House price impacts of racial, income, education, and age neighborhood segregation. J. Reg. Sci. **55**(3), 442–467 (2015)
5. Calmasur, G.: Determining factors affecting housing prices in turkey with hedonic pricing model. In: International Conference on Business and Economics Studies, Washington DC, USA, pp. 255–269 (2016)
6. Chong, T.T.L., Li, X.: The development of Hong Kong housing market: past, present and future. Econ. Political Stud., 1–20 (2019)
7. Duca, J.V., Muellbauer, J., Murphy, A.: Housing markets and the financial crisis of 2007–2009: lessons for the future. J. Financ. Stab. **6**(4), 203–217 (2010)
8. Eichholtz, P., Lindenthal, T.: Demographics, human capital, and the demand for housing. J. Hous. Econ. **26**, 19–32 (2014)
9. Eurostat: Housing in europe — 2024 edition. https://ec.europa.eu/eurostat/web/interactive-publications/housing-2024 (2024), interactive publication
10. Frayne, C., European Commission, D.G.f.E., Affairs, F.: Housing market developments in the euro area: focus on housing affordability. Tech. rep., European Commission (DG ECFIN) (2022). https://economy-finance.ec.europa.eu/system/files/2022-09/dp171_en.pdf
11. Fuller, G.W., Johnston, A., Regan, A.: Housing prices and wealth inequality in western Europe. West Eur. Polit. **43**(2), 297–320 (2020)
12. Goracy, D., Maciejewska, A., Maciuk, K.: The purchasing potential of EU residents in the real estate market in the context of sustained development. Sustainability **16**(23), 10373 (2024). https://doi.org/10.3390/su162310373, https://www.mdpi.com/2071-1050/16/23/10373
13. Guerreiro, A.E.R.: Determinantes no preço de um imóvel: uma análise hedónica no contexto algarvio. Master's thesis (2012)
14. Johnes, G., Hyclak, T.: House prices and regional labor markets. Ann. Reg. Sci. **33**(1), 33–49 (1999)
15. Katz, L., Rosen, K.T.: The interjurisdictional effects of growth controls on housing prices. J. Law Econ. **30**(1), 149–160 (1987)
16. Kaya, A., Atan, M.: Determination of the factors that affect house prices in turkey by using hedonic pricing model. J. Bus. Econ. Finance **3**(3), 313–327 (2014)
17. Lourenço, R.F., Rodrigues, P.M.M.: House prices in Portugal - what happened since the crisis? Tech. rep, Banco de Portugal, Economics and Research Department (2017)
18. McGreal, S., Taltavull de La Paz, P.: Implicit house prices: variation over time and space in Spain. Urban Stud. **50**(10), 2024–2043 (2013)

19. Moreira, M.S.G.: A dinâmica pública local e o valor da habitação: uma aplicação à área metropolitana do porto. repositorio-aberto.up.pt (2000)
20. Rappaport, J.: Consumption amenities and city population density. Reg. Sci. Urban Econ. **38**(6), 533–552 (2008)
21. Reuters: Portugal's property deals hit record, adding to shortage of affordable homes (2024). https://www.reuters.com/markets/europe/portugals-property-deals-hit-record-adding-shortage-affordable-homes-2024-12-23/
22. Tsatsaronis, K., Zhu, H.: What drives housing price dynamics: cross-country evidence. BIS Quart. Rev. (2004)
23. Varão, M.M.G.F.P.: The determinants of housing prices in Portugal between 2000 and 2015. Ph.D. thesis, Universidade Católica Portuguesa (2019)
24. Zhou, Q., Shao, Q., Zhang, X., Chen, J.: Do housing prices promote total factor productivity? Evidence from spatial panel data models in explaining the mediating role of population density. Land Use Policy **91**, 104410 (2020). https://doi.org/10.1016/j.landusepol.2019.104410

Modeling Risk in HIV Care: A Data-Driven Approach to Patient Profiling

M. Filomena Teodoro[1,2(✉)] and Teresa Oliveira[3,4]

[1] CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, 2810-001 Almada, Portugal

[2] CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, 1048-001 Lisboa, Portugal
mteodoro64@gmail.com, maria.teodoro@tecnico.ulisboa.pt

[3] CEAUL, Faculdade de Ciências, Universidade de Lisboa, 1649-014 Lisboa, Portugal
Teresa.Oliveira@uab.pt

[4] Departamento de Ciências e Tecnologia, Universidade Aberta, 1269-001 Lisboa, Portugal

Abstract. A statistical framework is developed to analyze and model the censored survival times of one hundred HIV-positive individuals. Censored data refers to observations where the exact value of the variable of interest (typically time until an event) is partially known. This occurs when the event (e.g., death, relapse, failure) has not yet happened for some individuals during the study period, or its occurrence is not observed for other reasons. As the observation period is limited, many lifetimes are right-censored, requiring appropriate statistical treatment. The analysis begins with the estimation of the survival function through nonparametric methods, providing an initial overview of T, the distribution of lifetime. In a second phase, the authors investigate how individual characteristics, such as age and type of medication, affect survival, employing the Cox proportional hazards model to quantify these relationships. The results contribute to a deeper understanding of the factors influencing life expectancy in HIV patients and offer valuable insights for public health planning.

Keywords: survival analysis · hazards · computational methods · AIDS

1 Introduction

The survival analysis has developed significantly since the last two decades of last century, mainly applied to medicine [1–5]. In literature, we can find a wide number of applications of survival models, namely in engineering [6], ecology [7], biometrics [8], physics [6], etc. For a first approach, we can find in [2,9,10] the basics of survival analysis techniques. More recent work, using Machine learning techniques, the authors of [11] explored deep Cox mixtures for survival regression. In [6] it is applied the censored expectation maximization algorithm for mixtures to intertrade waiting times. Also, the authors of [12] talk about interference and sensitivity analysis. In [13], the authors outcome risk model development for heterogeneity of treatment effect analysis, performing a comparison of non-parametric machine learning methods and semi-parametric

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 99–114, 2027.
https://doi.org/10.1007/978-3-032-30521-3_7

statistical methods. In the study proposed in [14], a comparison of the Cox model with the Fine-Gray model to take competing risks into account for survival analyzes of re-fracture rates is performed.

To evaluate some intervention strategies, such as predicting outbreaks or understanding avian influenza dynamics, Huang and their collaborators have searched and organized literature [15] classifying mathematical, statistical, and machine learning-based models as a tool to play a crucial role in risk assessment, understand transmission and predict outbreak, allowing to reduce the mortality index of birds and mamals, to improve the food security, reduce the rate transmission, and so on. In [16], a systematic review of statistical methods to evaluate the performance of survival prediction models from conventional to deep learning approaches.

The main objective of this work is to model the life time T of a group of one hundred of individuals with AIDS. The provided data, like most cases in survival analysis, are censored (right censored) since the study is usually carried out over a fixed period of time (or a fixed number of deaths is reached).

This work is divided in five sections. An introduction starts the manuscript; Sect. 2 details some aspects about survival analysis. Follows the description of the applied methodology in Sect. 3. In Sect. 4 some issues about the empirical application are described, starting with the basic statistical techniques, including the evaluation of some tests about the normality of data, followed by the construction and interpretation of the results. In Sect. 5 takes place some discussion and it is issued some final remarks.

2 Survival Analysis Some Concepts

In the present work, the author intend to determine and investigate the lifetime (T) distribution of one hundred individuals with HIV. The available data, similarly with the majority of cases in survival analysis, is censoured (on right) once the study is usually performed during a certain period or until there exists a certain number of deaths.

2.1 Survival Function

In this section we present some basic concepts of survival analysis [4,5,7,17–19].

A survival function $S(t)$ is defined as the individual surviving at an instant $t > 0$, and it is being given by

$$S(t) = 1 - F(t) = P(T > t),\ t > 0.$$

The failure rate function or hazard, $h(t)$, is defined as the probability of an individual dying in the next instant knowing that he survived until $t > 0$, and it is given by:

$$h(t) = \frac{f(t)}{S(t)}.$$

The cumulative hazard function $H(t)$ is defined as

$$H(t) = \int_0^t h(t)dt = -\log S(t).$$

It is important to understand whether the data to be studied correspond to moments of death or censorship, given that can be considered different approaches. At an early stage it is important to sketch some estimate of the survival function. Only after a qualitative analysis should models be estimated according to Cox (proportional hazards) or others.

2.2 Survival Function of Kaplan-Meier

The estimate of the nonparametric survival function, for censored data, called Kaplan-Meier, is given by

$$\hat{S}(t) = \prod_{i:t_i<t} \left(\frac{n_i - d_i}{n_i} \right),$$

with

- n_i the number individuals in risk immediately before t_i,
- d_i the number dead in t_i,
- l_i the number individuals censoured in $[t_i, t_{i+1}[$,

being $n_i = n - (d_1 + \cdots + d_{i-1}) - (l_0 + \cdots + l_{i-1})$.

One can also construct confidence intervals for $S(t)$ knowing the estimate of its variance.

$$\hat{S}(t) \pm 1.96\sqrt{\widehat{var}\left(\hat{S}(t)\right)}.$$

In this work, the confidence intervals obtained in some survival function plots used the Greenwood estimate for the variance.

2.3 Cox Proportional Hazard Models

The Cox proportional hazards model is one of the classes of regression models. These are models with proportional failure rate (hazard proportional). In these models it is assumed that given two individuals, with associated covariate vectors $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})$, with $i = 1, 2$, the ratio $h(t, \mathbf{z}_1)/h(t, \mathbf{z}_2)$ does not depend on t. With this premise the failure rate function T, given z, can be written in the form

$$h(t, \mathbf{z}) = h_0(t) g(\mathbf{z}).$$

A possible structure for $g(\mathbf{z})$ is given by

$$g(\mathbf{z}) = \exp(\beta^T \mathbf{z})$$

ensuring positivity. Note that the vector β^T measures the effect of the covariates. Each coefficient β_i associated with each covariate is often called the risk score associated with the variable z_i of $\mathbf{z}$. One question that can be asked is: what is the structure of the survival function for this model? It can be deduced obtaining:

$$S(t,\mathbf{z}) = \underbrace{S_0(t)}_{\text{associated to } h_0(t)}^{g(\mathbf{z})}$$

Coming back to the failure rate, how to estimate it? In the Cox model, it is non-parametric, obtaining a partially parametric model due to the influence of the covariates. Cox found a function that behaves like the likelihood function, not depending on the nonparametric $h_0((t)$. This function is given by

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\left(\mathbf{z}_i^T \beta\right)}{\sum_{l \in R_i} \exp\left(\mathbf{z}_l^T \beta\right)},$$

where

- $R_i = R_I(t)$ is the set of indices of individuals at risk at instant t,
- k the number of deaths in $t_1 < \cdots < t_k$,
- n the number of individuals under study.

A generalization of the Cox model is to admit the existence of non-proportional hazards between strata $h(t,\mathbf{z}) = h(t_i)g(\mathbf{z})$.

As with any regression model, we have all the inherent model validation steps: variable selection, model comparison, residual analysis (identification of outliers and influential observations), verification of Cox proportional hazards hypotheses, etc. Note that the tests used in linear models are again used in this context, highlighting t-ratios, likelihood ratio, etc.

Since we are estimating several survival functions, it is important to mention the existence of comparison tests between the different functions. The test used for this purpose is the Mantel-Haenzel-Savage (log-rank) test which, despite adding non-independent frequency tables, is quite powerful when the failure rate functions are approximately proportional.

3 Empirical Approach

3.1 Data

The data used in this article are reported to the first decade of this century, concern AIDS patients, a total of 100, and include the following variables:

- **ID**- identification code,
- **Time** - survival time (in months),
- **Age** - age in years,
- **Drug** - intravenous use (0 = “no”, 1 = “yes”),
- **Status** - 1 = “death due to AIDS”, 0 = “Obs. Censored”.

3.2 Survival Function of Kaplan-Meier

In the initial step we obtained the Kaplan-Meier estimate considering the censoured data, using the SPLUS software.

```
> KM<-survfit(Surv(Tempo, Status), conf.type="plain",
 type="kaplan-meier",data = Dados2)
> plot(KM)
```

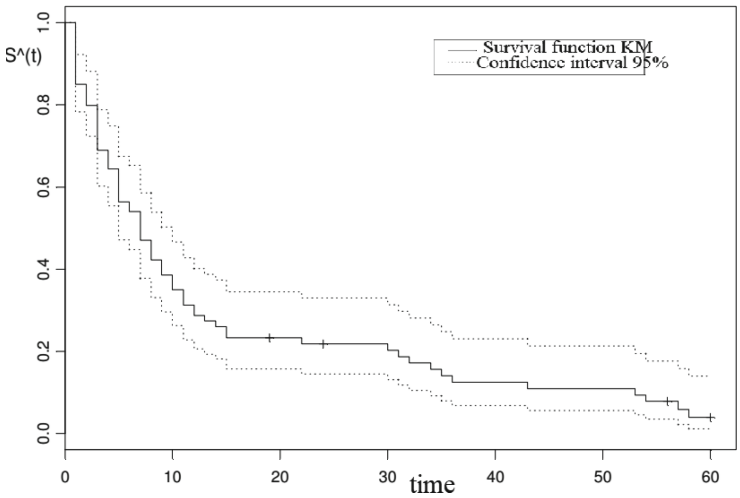


Fig. 1. Empirical Survival graph. Kaplan-Meier estimate.

Figure 1 represents the empirical survival curve using the Kaplan-Meier estimate due the fact that it is decreasing with concavity facing upwards it is similar to the survival function of the exponential distribution.

A more detailed analysis can be done by determining the Kaplan-Meier estimate, disaggregating by type of drug (intravenous - no, intravenous - yes).

```
  >KMdroga<-survfit(Surv(Tempo, Status) ~as.factor(Droga),conf.type="plain"
 ,type="kaplan-meier",data = Dados2)
 > plot(KMdroga, lty=2:3, xlab="T", ylab="Prob.  de sobrev." )
 > legend(30, 0.80, legend=c(" intra-venosa-nao"," intra-venosa-sim"
), lty=2:3)
```

Figure 2 indicates distinct survival empirical functions per kind of drug. Besides both curves be similar to exponential distribution (concavity facing upward), the curve associate to "*no intravenous* drug evidences a longest time of survival. This fact evidences that drug kind hall be considered explanatory variable in the model. Other variables in the data were age and gender, they were also considered as covariates in the model to estimate.

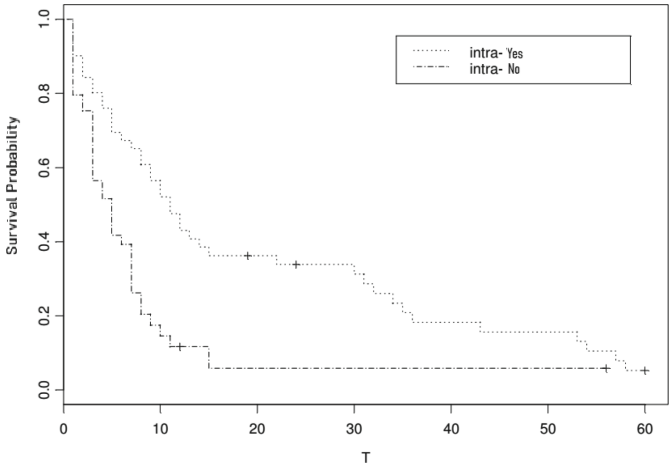


Fig. 2. Empirical Survival graph per level of drug (intravenous - yes, intravenous - no). Kaplan-Meier estimate.

3.3 Model 1

It was estimated the model of Cox Proportional Hazards, including all possible information as covariates, the type of drug and age respectively. The output reveals good indicators, the "*t-ratios*" associated to covariates coefficients are significant, meaning that this explanatory variables shall be considered. Other indication in the same way are the confidence intervals obtained for the exponential of each coefficient: none contains the value one. The confidence interval [20] for the age is farm from of 1:

```
      *** Cox Proportional Hazards ***
Call:
coxph(formula = Surv(Tempo, Status) ~ Idade + Droga, data = Dados2)

  n= 100

        coef exp(coef) se(coef)     z        p
Idade 0.0971      1.10   0.0186 5.21 1.9e-007
Droga 1.0167      2.76   0.2562 3.97 7.2e-005

      exp(coef) exp(-coef) lower .95 upper .95
Idade      1.10      0.907      1.06      1.14
Droga      2.76      0.362      1.67      4.57

Rsquare= 0.324   (max possible= 0.997 )
Likelihood ratio test= 39.1  on 2 df,   p=3.18e-009
Wald test            = 36.1  on 2 df,   p=1.43e-008
Score (logrank) test = 38.4  on 2 df,   p=4.6e-009
```

$$\begin{aligned} \exp(\beta_{\text{age}}) &: [1.06, 1.14], \\ \exp(\beta_{\text{drug}}) &: [1.67, 4.65]. \end{aligned}$$

The tests that allow to compare the estimated model and the model with explanatory variables, namely the likelihood ratio test, the wald test and the logrank test lead to largely significant statistical values.

The estimate of the survival function (model 1) is given by the following code:

```
> summary(survfit(cox.fit1))
Call: survfit.coxph(object = cox.fit1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    100      15 0.885870 0.02969    0.8295436       0.9460
    2     83       5 0.842263 0.03518    0.7760556       0.9141
    3     73      10 0.740745 0.04557    0.6566098       0.8357
    4     61       4 0.692477 0.04944    0.6020504       0.7965
    5     56       7 0.597255 0.05501    0.4986109       0.7154
    6     49       2 0.567512 0.05620    0.4673935       0.6891
    7     46       6 0.472812 0.05854    0.3709334       0.6027
    8     39       4 0.401388 0.05911    0.3007497       0.5357
    9     35       3 0.349514 0.05830    0.2520481       0.4847
   10     32       3 0.300375 0.05642    0.2078603       0.4341
   11     28       3 0.252605 0.05370    0.1665260       0.3832
   12     25       2 0.220606 0.05144    0.1396798       0.3484
   13     21       1 0.200386 0.05060    0.1221582       0.3287
   14     20       1 0.180206 0.04949    0.1051961       0.3087
   15     19       2 0.143205 0.04612    0.0761761       0.2692
   22     16       1 0.124105 0.04412    0.0618267       0.2491
   30     14       1 0.105184 0.04169    0.0483721       0.2287
   31     13       1 0.088282 0.03870    0.0373887       0.2085
   32     12       1 0.072868 0.03531    0.0281872       0.1884
   34     11       1 0.058942 0.03160    0.0206068       0.1686
   35     10       1 0.046215 0.02762    0.0143236       0.1491
   36      9       1 0.035166 0.02347    0.0095066       0.1301
   43      8       1 0.023454 0.01880    0.0048757       0.1128
   53      7       1 0.015072 0.01416    0.0023902       0.0950
   54      6       1 0.008552 0.00968    0.0009294       0.0787
   57      4       1 0.003895 0.00557    0.0002364       0.0642
   58      3       1 0.000872 0.00192    0.0000117       0.0649
```

The graphic of the survival function (model 1) is displayed in Fig. 3.

3.4 Model 2

The second model considers only one explanatory variable, the age.

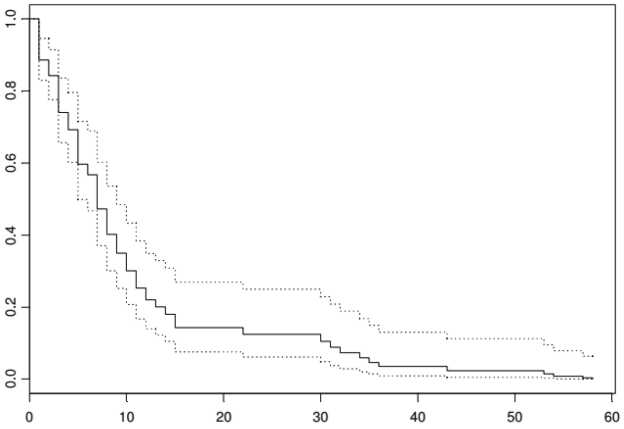


Fig. 3. Model 1-Cox Proportional Hazards. Explanatory variables. Age, Drug type.

```
> summary(survfit(cox.fit1))
Call: survfit.coxph(object = cox.fit1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    100      15 0.885870 0.02969    0.8295436       0.9460
    2     83       5 0.842263 0.03518    0.7760556       0.9141
    3     73      10 0.740745 0.04557    0.6566098       0.8357
    4     61       4 0.692477 0.04944    0.6020504       0.7965
    5     56       7 0.597255 0.05501    0.4986109       0.7154
    6     49       2 0.567512 0.05620    0.4673935       0.6891
    7     46       6 0.472812 0.05854    0.3709334       0.6027
    8     39       4 0.401388 0.05911    0.3007497       0.5357
    9     35       3 0.349514 0.05830    0.2520481       0.4847
   10     32       3 0.300375 0.05642    0.2078603       0.4341
   11     28       3 0.252605 0.05370    0.1665260       0.3832
   12     25       2 0.220606 0.05144    0.1396798       0.3484
   13     21       1 0.200386 0.05060    0.1221582       0.3287
   14     20       1 0.180206 0.04949    0.1051961       0.3087
   15     19       2 0.143205 0.04612    0.0761761       0.2692
   22     16       1 0.124105 0.04412    0.0618267       0.2491
   30     14       1 0.105184 0.04169    0.0483721       0.2287
   31     13       1 0.088282 0.03870    0.0373887       0.2085
   32     12       1 0.072868 0.03531    0.0281872       0.1884
   34     11       1 0.058942 0.03160    0.0206068       0.1686
   35     10       1 0.046215 0.02762    0.0143236       0.1491
   36      9       1 0.035166 0.02347    0.0095066       0.1301
   43      8       1 0.023454 0.01880    0.0048757       0.1128
   53      7       1 0.015072 0.01416    0.0023902       0.0950
   54      6       1 0.008552 0.00968    0.0009294       0.0787
   57      4       1 0.003895 0.00557    0.0002364       0.0642
   58      3       1 0.000872 0.00192    0.0000117       0.0649
```

All tests were largely significant. A last test to perform is the comparison this model with model 1. To test the best model, we simply consider the null hypothesis the coefficient of the drug variable as null against all other alternatives. By the likelihood ratio

we have that

$$\Lambda_{1,2|H_0} = -2\left(\ln L\left(\widetilde{\beta}\right)_{\text{mod }2} - \ln L\left(\widehat{\beta}\right)_{\text{mod }1}\right) \overset{a}{\sim} \chi^2_{(1)}$$

The null hypothesis is rejected when

$$\Lambda_{1,2obs} > \chi^2_{(1,1-\alpha)}.$$

When comparing the two models we have There is a significant difference between the two models as expected from the analysis already done in model 1. Another model was studied. It only presented the drug variable as an explanatory variable. The conclusions are similar: model 1 should be considered.

$$\Lambda_{1,2obs} = 39.1 - 23.3 = 15.8$$
$$p - value = 1 - \chi^2_{(15.8,1)} = 7.0403 \times 10^{-5}.$$

3.5 Model 3

The next step was to estimate models for each level of the drug variable, that is, to assume the existence of a stratification factor that allows the existence of non-proportional hazards between the strata. The obtained model (model 3) was

```
> cox.2<-coxph(Surv(Tempo, Status)~Idade+strata(Droga), data=Dados2)
> summary(cox.2)
Call:
coxph(formula = Surv(Tempo, Status) ~ Idade + strata(Droga), data = Dados2)

  n= 100

        coef exp(coef) se(coef)    z        p
Idade 0.0923       1.1   0.0188 4.91 9.2e-007

      exp(coef) exp(-coef) lower .95 upper .95
Idade       1.1      0.912      1.06      1.14

Square= 0.216   (max possible= 0.992 )
Likelihood ratio test= 24.3  on 1 df,   p=8.22e-007
Wald test            = 24.1  on 1 df,   p=9.2e-007
Score (logrank) test = 24.5  on 1 df,   p=7.41e-007
```

In this model, like before, all tests are significant. To compare the survival curves in the two strata, we used the logrank test. Since the test is largely significant, there is a clear difference in survival functions between strata. There is also statistical evidence that the age variable is statistically significant.

```
survdiff(formula = Surv(Tempo, Status) ~ Idade + strata(Droga), data = Dados2)

         N Observed Expected (O-E)^2/E (O-E)^2/V
Idade=20 1        0    2.409   2.40947   3.30534
Idade=21 1        1    2.142   0.60853   0.67873
Idade=22 1        1    2.725   1.09190   1.27708
Idade=25 2        1    4.500   2.72206   3.29818
Idade=26 2        2    1.914   0.00391   0.00498
Idade=28 3        2    1.089   0.76191   0.85593
Idade=29 2        1    3.644   1.91861   2.24068
Idade=30 6        4    4.847   0.14813   0.18045
Idade=31 5        4    5.666   0.48993   0.67233
Idade=32 9        8    6.238   0.49767   0.65192
Idade=33 6        5    5.473   0.04083   0.04928
Idade=34 8        8    7.292   0.06878   0.08471
Idade=35 5        2    3.523   0.65842   0.80318
Idade=36 9        8   10.342   0.53050   0.68169
Idade=37 3        2    0.830   1.64923   1.86199
Idade=38 3        3    0.769   6.46638   7.83106
Idade=39 5        5    3.923   0.29546   0.38819
Idade=40 3        2    1.457   0.20240   0.25347
Idade=41 4        3    2.344   0.18386   0.22059
Idade=42 4        3    1.842   0.72812   0.89297
Idade=43 2        2    1.795   0.02334   0.02552
Idade=44 4        2    1.080   0.78478   0.99088
Idade=45 1        1    0.508   0.47611   0.60322
Idade=46 3        2    1.340   0.32503   0.39086
Idade=47 4        4    1.464   4.39550   4.93894
Idade=48 1        1    0.212   2.92819   3.17114
Idade=50 1        1    0.204   3.10408   3.90000
Idade=51 1        1    0.163   4.28858   4.68661
Idade=54 1        1    0.265   2.04298   2.20111

 Chisq= 54.7  on 28 degrees of freedom, p= 0.00183
```

```
> survdiff(Surv(Tempo, Status)~Droga, data=Dados2)
Call:
survdiff(formula = Surv(Tempo, Status) ~ Droga, data = Dados2)

          N Observed Expected (O-E)^2/E (O-E)^2/V
Droga=0  51       42     54.9      3.02      11.9
Droga=1  49       38     25.1      6.60      11.9

 Chisq= 11.9  on 1 degrees of freedom, p= 0.000575
```

4 Analysis of Residuals

The residuals were analyzed for both model 1 and model 3. Since these analyzes are very similar, this article only focuses on the analysis of model 3.

In the first phase, the covariate age was investigated to ensure it contributed in the correct functional form. Graphical analysis of the residues is essential. We can find in next figures some plots with interest.

From observation of Fig. 4, we tried models where the covariate age had various shapes close to line in Fig. 4: square root, polygonal line, polynomial shape. The best models obtained, despite being more complex, did not have results that were significantly different from model 3.

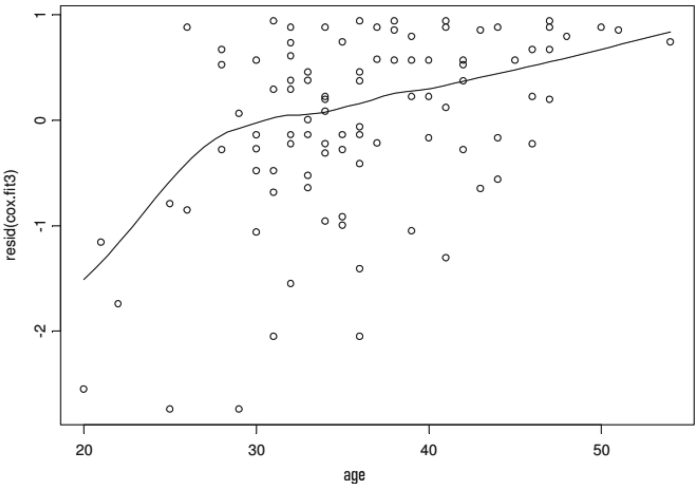


Fig. 4. Model 3. Residuals versus Age.

Also, no poorly adjusted individuals were identified (seen Fig. 5). In identifying influential observations, "dfbetas" plots were made, which give us the difference in the estimate when observation is omitted (see Fig. 6). Notice that an observation is considered influential if the absolute value of this difference is greater than 0.5.

To verify the proportional hazards assumption, a graphical analysis of the estimated values for the coefficients of the covariates is important: whether they can be considered constant or not. As many graphs are made as there are explanatory variables considered. In model 3 there is only one covariate (age), we only need to analyze a single graph. Figure 7 displays the estimates of coefficient beta of age along the time.

Estimates of the beta coefficient over time lead to confidence bands, which, through graphical analysis, appear to limit a horizontal line, of the type $\beta(t) = \beta$. The relative linearity test for beta consists of investigating whether the β estimates are correlated with time. The following output was obtained:

```
> cox.zph(cox.2)
          rho  chisq     p
Idade -0.0176 0.0223 0.881
>
```

The test of null correlation with the time conduced to $pvalue = 0.881$, meaning that there exists evidence to not reject the horizontal linearity. The model 3 (model the Cox is not rejected and we obtain the following curves of survival. By the observation of Fig. 8, it is clear that the survival curve relative to the stratum of individuals who consume intravenous drugs is lower.

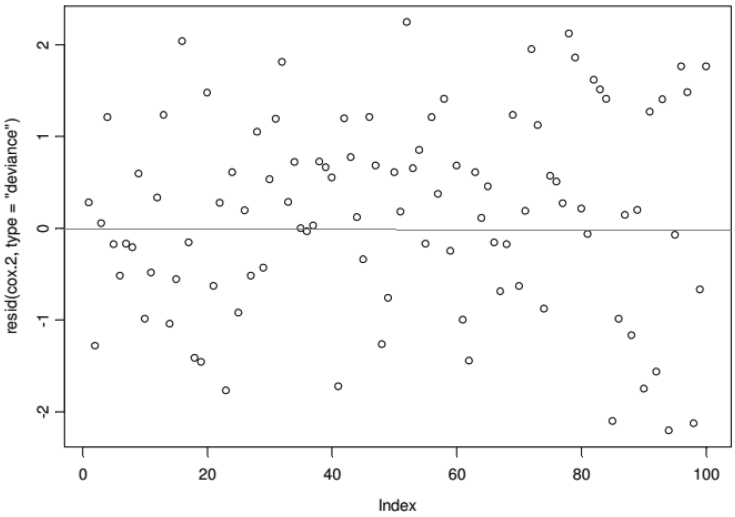


Fig. 5. Model 3. Residuals versus index.

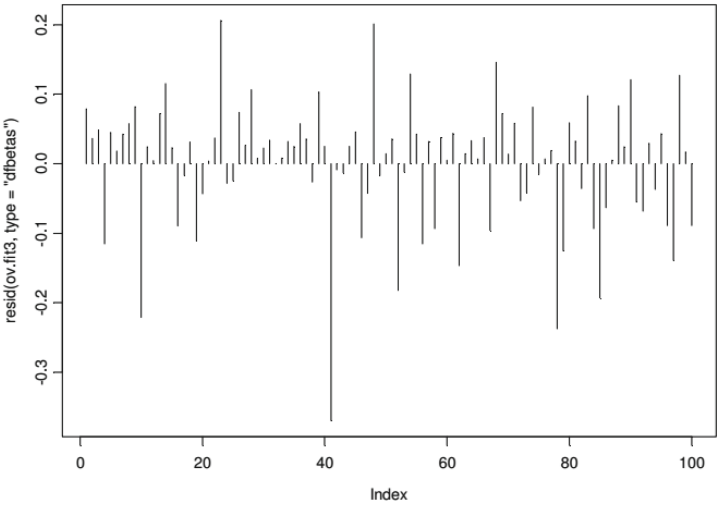


Fig. 6. Model 3. DFbetas versus index.

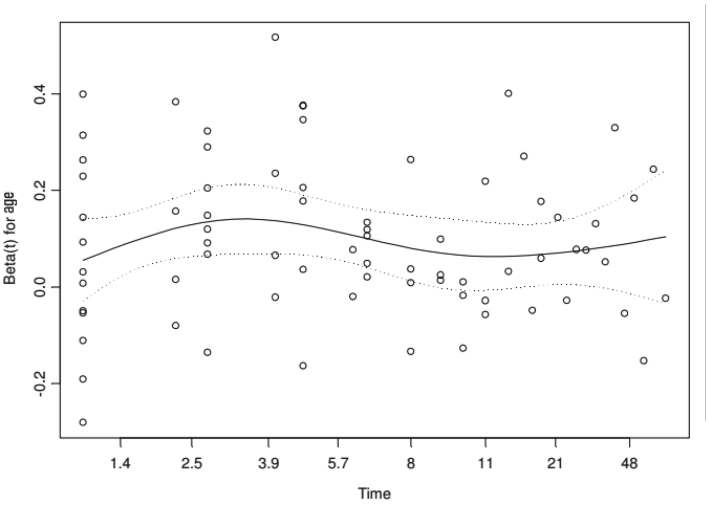


Fig. 7. Model 3. Estimates of coefficient β along the time.

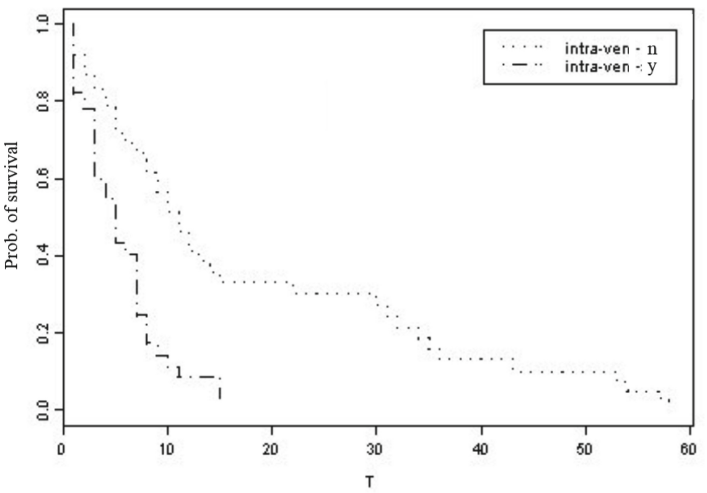


Fig. 8. Model 3. Curves of Cox survival function per drug strata.

5 Discussion and Final Remarks

Model 3 is a hazard proportional by strata depending on the drug values. The estimated model comes in the form

$$\widehat{h}(t,\ \text{idade}) = \widehat{h}_{droga}(t)\exp(\widehat{\beta} \times \text{idade}),\ \text{so}$$

$$\exp(\widehat{\beta} \times \text{idade}) = \exp(0.0923 \times \text{idade}).$$

It can be concluded that the risk factor increases $exp(0.0923) = 1.1$ when age increases by 1 year, at each drug level. The 95% risk confidence interval for the age is $(1.06, 1.14)$. The survival functions obtained accordingly with model 3 clearly confirm this fact. Survival of individuals taking intravenous drugs is lower. This behavior is identical to what we already had in the empirical survival functions of Kaplan-Meier. It is clear that age is an important explanatory factor in the survival of the individual.

In model 1 the model is a hazard proportional without stratification. The estimate of the model is given by

$$\begin{aligned}\widehat{h}(t,\ \text{idade,droga}) &= \widehat{h}(t)\exp(\widehat{\beta}_1 \times \text{idade} + \widehat{\beta}_2 \times \text{droga}) \\ &= \widehat{h}(t)\exp(\widehat{\beta}_1 \times \text{idade})\exp\left(\widehat{\beta}_2 \times \text{droga}\right)\end{aligned}$$

onde

$$\begin{aligned}&\exp(\widehat{\beta}_1 \times \text{idade})\exp\left(\widehat{\beta}_2 \times \text{droga}\right) \\ &= \exp(0.0971 \times \text{idade})\exp(1.0167 \times \text{droga}).\end{aligned}$$

In model 1 we can conclude that the risk factor increases $exp(0.0971) = 1.1$ when the age increases by 1 year, when the drug variable remains the same. Note that the risk interval obtained at 95% does not contain the value 1, $(1.06, 1.14)$, despite being very close at its left end. Age is an important explanatory variable. Regarding the drug variable, if this changes from 0 to 1, the group of individuals who use intravenous drugs has a higher risk factor given by $exp(1.10167) = 2.76$. The relative risk interval obtained at 95% does not contain the value 1 and has an extreme well above the interval obtained for age: $(1.67, 4.57)$.

We highlight that both models agree on their conclusions: age is an important variable for an individual's survival, the kind of drug is an important factor for survival. In both models this explanatoru variable was highly significant. In future work, we intend to explore the same data, but knowing that database was limited, the next step would be to add more information such as gender, some familial diseases, etc. We would also like to apply other techniques that have emerged more recently, which are extensions of the methods used in this article [11, 13, 14, 21–23].

Acknowledgements. The first author was supported by Portuguese funds through the *Center of Naval Research* (CINAV), Naval Academy, Portugal and *The Portuguese Foundation for Science and Technology* (FCT), through the *Center for Computational and Stochastic Mathematics* (CEMAT), University of Lisbon, Portugal, project DOI: (https://doi.org/10.54499/UID/04621/2025).

References

1. Gill, R.D.: Understanding cox's regression model: a martingale approach. J. Am. Stat. Assoc. **79**(386), 441–447 (1984). http://www.jstor.org/stable/2288288. Accessed 7 May 2026
2. Gill, R.D.: Lectures on Survival Analysis, pp. 115–241. Springer, Heidelberg (1994). https://doi.org/10.1007/BFb0073873. Accessed 9 May 2026
3. Lee, E., Go, O.: Survival analysis in public health research. One Health **18**, 105–134 (1997). https://doi.org/10.1146/annurev.publhealth.18.1.105. Accessed 9 May 2026
4. Harrell, F., Lee, K., Mark, D.: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. **15**, 361–387 (1996)
5. Christensen, E.: Multivariate survival analysis using cox's regression model. Hepatology **7**, 1346–1358 (1987)
6. Kreer, M., Kizilersu, A., Thomas, A.W.: Censored expectation maximization algorithm for mixtures: application to intertrade waiting times. Physica A: Stat. Mech. Appl. **587**, 126456 (2022). https://www.sciencedirect.com/science/article/pii/S0378437121007299. Accessed 9 May 2026
7. Kenneth, H.P., Scott, R.W., Christine, M.B., Paul, D.C.: Survival analysis in telemetry studies: the staggered entry design. J. Wildlife Manage. **53**(1), 7–15 (1989). http://www.jstor.org/stable/3801296
8. David, A.S.: Sample-size formula for the proportional-hazards regression model. Biometrics **39**(2), 499–503 (1983). http://www.jstor.org/stable/2531021
9. Lawless, J.: Statistical Models and Methods for Lifetime Data, 2nd edn. Wiley, New York (2003)
10. Rocha, C., Papoila, A.L.: Análise de Sobrevivência. Sociedade Portuguesa de Estatística., Lisboa (2009), https://www.spestatistica.pt/storage/app/uploads/public/600/481/dc3/600481dc32725998871922.pdf. Accessed 9 May 2026
11. Nagpal, C., Yadlowsky, S., Rostamzadeh, N., Heller, K.: Deep cox mixtures for survival regression. In: Simos, T., et al. (eds.) Proceedings of Machine Learning Research, vol. 126, pp. 1–27. AIP (2021, in press)
12. Tyler, J.V., Eric, J.T.T., Elizabeth, H.M.: Interference and sensitivity analysis. Stat. Sci. **29**(4), 687–706 (2014)
13. Xu, E., et al.: Outcome risk model development for heterogeneity of treatment effect analyses: a comparison of non-parametric machine learning methods and semi-parametric statistical methods. BMC Med. Res. Methodol. **24**(1), 1–9 (2024). https://doi.org/10.1186/s12874-024-02265-8
14. Nolan, E.K., Chen, H.-Y.: A comparison of the Cox model to the Fine-Gray model for survival analyses of re-fracture rates. Arch. Osteoporos. **15**(1), 1–8 (2020). https://doi.org/10.1007/s11657-020-00748-x
15. Huang, S., et al.: A systematic review of mathematical and machine learning models of avian influenza. One Health **21**, 2109–2123 (2025). https://www.sciencedirect.com/science/article/pii/S2352771425002393. Accessed 9 May 2026

16. Park, S., Park, J., Kim, H., Park, S.: Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches). Korean J. Radiol. **22**(10), 1697–1707 (2021). https://doi.org/10.3348/kjr.2021.0223. Accessed May 9th 2026
17. Rosner, B.: Fundamentals of Biostatistics, 6th edn. Pacific Grove, Duxbury (2006)
18. Long, J.: Regression Models for Categorical and Limited Dependent Variables. Sage Publications, Thousand Oaks (1987)
19. Peduzzi, P., Concato, J., Feinstein, A., Holford, T.: Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J. Clin. Epidemiol. **48**, 1503–1510 (2004)
20. Pencina, M., D'Agostino, R.: Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat. Med. **23**(23), 2109–2123 (2004)
21. Goerdten, J., Carrière, I., Muniz-Terrera, G.: Comparison of cox proportional hazards regression and generalized cox regression models applied in dementia risk prediction. Alzheimers Dement **6**(1), e12041 (2020). http://www.jstor.org/stable/2288288. Accessed 9 May 2026
22. Ahmed, H.S.: Advanced statistical methods for hazard modeling in cardiothoracic surgery: a comprehensive review of techniques and approaches. Indian J. Thoracic Cardiovascular Surgery **40**(5), 633–644 (2020). http://www.jstor.org/stable/2288288. Accessed 9 May 2026
23. Hélio, A.M.: Survival analysis in clinical and experimental studies. J. Vasc. Bras. **16**(4), 267–269 (2017). https://doi.org/10.1590/1677-5449.001604. Accessed 8 May 2026

The Influence of Fatigue on Decision-Making in Military Context

Francisco Abreu[1], Nuno Almeida[2], Renato Santos[3], Rui Lucena[2], and Paula Simões[4(✉)]

[1] Portuguese Military Academy, Lisboa, Portugal
abreu.fjg@exercito.pt

[2] Military Academy Research Centre - Military University Institute(CINAMIL) and MRLab-Military Readiness Lab-Portugal and CIPER – Faculdade de Motricidade Humana, Universidade de Lisboa, Lisboa, Portugal
almeida.nrc@academiamilitar.pt, lucena.rjp@exercito.pt

[3] Military Academy Research Centre - Military University Institute (CINAMIL) and Higher Institute of Intercultural and Transdisciplinary Studies of Almada, University Campus of Almada, Almada, Portugal
santos.recp@exercito.pt

[4] Military Academy Research Centre - Military University Institute (CINAMIL) and MRLab-Military Readiness Lab-Portugal and NOVA MATH - Centre for Mathematics and Applications, NOVA University of Lisbon and Engineering Superior Institute of Lisbon (ISEL), Polytechnic Institute of Lisbon, Lisboa, Portugal
paula.simoes@academiamilitar.pt

Abstract. Decision making is fundamental in military operations, requiring speed and accuracy in increasingly uncertain scenarios. In such contexts, military personnel face high physical and cognitive demands that can cause fatigue and affect their operational performance. This study analysed the influence of fatigue on decision making in a military context, aiming to optimize performance in demanding situations. A sample of 4th year students from the Portuguese Military Academy was selected. Each participant completed the RP 30 Problem Solving Test, which assesses cognitive abilities essential to decision making, on two different measurement moments. Firstly, the individuals performed a ruck march carrying an approximate load of 30 kg, second, on another day, they completed the test in a resting condition. The data collected on the physiological parameters in the participants, namely heart rate, continuously monitored during the march, enabling calculation of internal load through Training Impulse and also the calculation of the Summated heart rate zone method, as well as changes in the percentage of heart rate reserve. The data was analysed employing various descriptive statistical methods, combined with adequate statistical inference techniques and probability models, using IBM SPSS Statistics 30 and Microsoft Excel software. Two indexes are proposed, the Decision-Making Speed Index and the Decision-Making Performance Index, which quantify, respectively, the speed and quality of decisions under fatigue providing standardized metrics for future studies. The indexes revealed that, when fatigued, cadets experienced a 23,33% reduction in decision-making speed and a 15,09% reduction in decision-making quality. The results demonstrated that fatigue significantly compromises both the speed and quality

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 115–131, 2027.
https://doi.org/10.1007/978-3-032-30521-3_8

of decision making, impairing several key cognitive abilities. It is concluded that implementing training protocols that combine intense physical effort with decision making tasks is fundamental for deepening understanding as an effective measure to enhance the functional integration of physical and cognitive capacities in order to optimize the operational performance of military personnel, even under exhaustive conditions.

Keywords: Decision Making · Fatigue · Physical Load · RP 30 · Statistical hypothesis tests

1 Introduction

Decision-making can be defined, in a simple and objective way, as the process of analysing several options and choosing the most appropriate one for a specific situation [1]. According to Deveci et al. (2020) [2], the decision-making process comprises the rigorous assessment of environmental factors, integrating a comprehensive perspective, the identification of problems and their potential solutions, promoting an organized analysis that supports the informed choice among the available alternatives.

1.1 Fatigue

Fatigue, as a multidimensional phenomenon, integrates both physical and mental dimensions. The physical dimension is manifested by the reduction of the ability to perform physical efforts. On the other hand, the mental dimension translates into difficulties in terms of attention and the execution of cognitive tasks [3].

The state of mental fatigue corresponds to the brain's response to certain specific behaviours that lead to fatigue as a result of routine repetition. In this type of fatigue, a prolonged specific action decreases cognitive activity, which can result in difficulty concentrating, memory lapses, irritability, and a general feeling of mental exhaustion. While mental fatigue is reflected in reduced ability to focus, loss of motivation, and emotional tiredness, it can also be reflected in reduced physical fitness [4].

The concept of physical fatigue consists of a temporary limitation of muscle capacity to sustain optimal levels of performance, arising as a consequence of prolonged efforts or excessive workloads, and leading to a momentary reduction in the ability to perform physical activities [5]. Physical fatigue can be classified into central and peripheral. Central fatigue corresponds to a limitation in sending signals by the central nervous system to activate muscles, while peripheral fatigue results from failures in the neuromuscular system, which prevent the sustainment of muscle contraction after receiving stimulation from the central system [6]. Additionally recent research has shown that the physical and mental dimensions of fatigue interact in a significant way. In high-demand scenarios, physical fatigue has a direct impact on mental fatigue, namely affecting reaction time in cognitive tasks [7].

1.2 Fatigue in Military Context

Extreme physical and mental demands are often placed on military personnel in their daily lives. The accumulation of these factors can compromise both the individual performance and the functioning of military units, which directly compromises operational effectiveness.

In this context, service members face prolonged periods of intense physical exertion, exposure to adverse environmental conditions, and sleep deprivation. These conditions contribute significantly to muscle fatigue and physical exhaustion in military personnel [8]. In addition, operational demands often impose periods of continuous surveillance and regular shift changes, which result in irregular and insufficient sleep patterns, favouring the development of chronic physical fatigue and compromising performance during operations [9]. Furthermore, fatigue has a crucial impact on the cognitive ability of military personnel thus, when subjected to fatigue, military personnel show a significant reduction in their ability to adapt their decisions to complex and dynamic tactical situations, which leads to a decline in cognitive performance after being exposed to operational stress [10]. The same authors noted that, despite several studies quantify the physical and cognitive decrease resulting from this stress, there is a lack of studies that explore how such changes directly influence decision-making in a military context.

Military personnel are exposed to intense cognitive challenges, such as strategic planning and decision-making under pressure. The impact of cognitive demands on military personnel during various types of foot operations can manifest itself through difficulty concentrating, reduced reasoning ability and increased operational errors, compromising the safety and effectiveness of missions[11]. In addition, the prolonged nature of these operations entails a high workload and frequent sleep deprivation that negatively affect the memory and attention of military personnel, increasing the risk of critical errors in operational environments. Thus, in order to preserve cognitive clarity during such operations it has been suggested that cognitive resilience is critical to preserve operational effectiveness during military operations [12].

It has been reported that intense physical exertion, sleep deprivation, caloric restriction and stress resulting from life-or-death decisions result in significant physiological and psychological changes in military personnel. This set of factors can aggravate the state of fatigue, compromising both the physical and cognitive capacities of service members in combat [13]. Military personnel often face extremely demanding circumstances, and the combination of these factors leads to a significant increase in fatigue and a decrease in operational performance [14].

It becomes essential to constantly monitor fatigue in military settings, as the interaction between physical and mental fatigue can compromise readiness, decision-making, and, consequently, the success of military operations.

1.3 Characterisation of Military Marches

During military operations, walking is a fundamental capability, being a common and physically demanding practice [15]. During training, or even in war scenarios, military personnel are required to move with equipment and supplies, over varied terrain and at different speeds, for prolonged periods [16]. Furthermore, the weight of load

and distances to be covered must be carefully considered in order to maximize the performance of the military in subsequent missions. According to Weyden et al. (2022) [17], this activity can induce not only physical fatigue, but also mental fatigue, and the higher load-to-bodyweight ratio, the more compromised the cognitive function of military personnel will be after marching.

Walking becomes a vital skill for the mobility of service members, preparing them to face adverse and unpredictable scenarios and ensuring the effectiveness of operations. In this context, mobility is an indispensable resource, capable of making a difference on the battlefield [18]. Marching with a load is an essential parameter to assess the tactical preparedness of military forces. However, there is a limited amount of research looking at the impact of these marches on the cognitive ability of military personnel, which could serve to constitute an effective method to monitor fatigue and cognitive performance in operational scenarios [19].

In this sense, fatigue develops progressively over the course of loaded walks, as a consequence of its prolonged and strenuous nature. Fatigue derived from load transport is responsible for significant changes in the physical performance, negatively influencing the mechanics of their movement. These changes increase susceptibility to injury, compromising operational capacity [20].

In addition, carrying loads combined with prolonged physical exertion, compromises the ability to inhibit automatic or impulsive responses in the military, an essential component of cognitive control. The combination of these conditions, observed during loaded gaits, exerts a substantial negative impact on cognitive functions, reflected by a decrease in the accuracy of responses and a reduction in sensitivity to discriminate relevant stimuli. In addition, the increase of the load carried, as well as the frequency of the gaits, intensifies these adverse effects, also increasing the percentage of the heart rate reserve (%HRR) and the rate of perceived exertion of the subjects [21]. Thus, it is possible to identify the existence of a relationship between physical overload and cognitive impairment [19].

1.4 Fatigue vs Decision Making - Decision Time and Quality

The decision-making process often occurs in demanding and complex contexts. In a military context, the physical exhaustion resulting from operations accumulates with cognitive demands, which can compromise the decision-making capacity of individuals. When an individual is cognitively fatigued, it takes longer to make decisions, especially in situations that require adaptation or change [22].

Fatigue has a direct impact on cognitive processing speed, affecting coordination and information processing [23]. Given that cognitive speed is an essential element for effective decisions under pressure, any reduction in this ability can compromise the speed of information analysis and the accuracy of responses in situations that require an immediate reaction. In addition, sleep deprivation or inadequate rest intensify this effect, prolonging reaction time and negatively affecting decision-making [24].

However, in situations of extreme fatigue, individuals tend to adopt compensatory strategies that allow the maintenance or even a slight increase in processing speed, even if this occurs at the expense of a deterioration in accuracy. This adaptation suggests an

attempt to mitigate the effects of fatigue by prioritizing speed, which may result in less rigorous assessments, increasing the risk of error in critical decisions [25].

1.5 Influence of Fatigue in the Decision-Making Process

In highly demanding contexts, the accuracy in the assessment of information and the selection of the most appropriate response can be impaired, increasing the risk of errors and less effective choices. Despite the impact on the speed of decision-making, fatigue also compromises the quality of decisions [26].

Mental fatigue exerts a negative influence on decision-making, leading individuals to adopt a more conservative approach, demonstrating greater risk aversion in the face of uncertainty or potentially risky options [27]. Fatigue can also reduce emotional control and the ability to assess risks, which results in a distorted perception of the risks and benefits associated with different alternatives, thus compromising the quality of the decision [28]. This phenomenon is characterized by a gradual reduction in motivation, particularly in contexts that require intense and continuous cognitive effort. This state compromises essential skills for operational effectiveness, such as the management of cognitive functions, negatively influencing the execution of complex tasks. Its effects are more pronounced on controlled cognitive processes, also known as top-down processes, which concern voluntary mental operations, such as planning, evaluation of alternatives, logical reasoning and conscious decision-making [29].

Recently, Brady et al. in [30] point out, fatigue can also impair decision-making by reducing attention and increasing the propensity for error. At the same time, it may lead to a decrease in cognitive flexibility and the ability to adapt to new information and scenarios [30]. Thus, it is essential to deepen the analysis of the influence of fatigue on decision-making in a military context, contributing to clarify the effects of this phenomenon on operational performance.

1.6 Document Structure

This work is structured as follows: Sect. 2 describes the objectives and the methodology applied; the data, the results analysis and discussion will be presented in Sect. 3, and, finally, conclusions and final remarks close the document in Sect. 4.

2 Objectives and Methods

2.1 Objectives

The main objective of the present research consists of "Analysing the influence of fatigue on decision-making in a military context, in order to optimize operational performance in demanding contexts".

Following this general objective, the hypothesis by research question: "What is the influence of fatigue on decision-making in a military environment?". This question guides the structuring of the research, configuring itself as the central question that sustains the analysis of the phenomenon under study.

Furthermore, we aim to understand the physiological impact of a loaded march on the military personnel and to what extent does fatigue interfere with the speed and quality of decision-making?

Participation in the study was preceded by obtaining a written informed consent, in accordance with the ethical principles of scientific research with human beings. In this context, the participants were presented with the objectives of the research, as well as a detailed explanation of the activities to be carried out, including what would be asked of them, the estimated time of participation and the place of data collection. Information was also provided regarding the processing of the data collected, namely the guarantee of confidentiality and anonymity, as well as the procedures for storing and preserving the information. Additionally, it was made explicit that participation was entirely voluntary, and participants could refuse or interrupt their collaboration at any time, without any prejudice or penalty.

2.2 Experimental Protocols

The Problem-Solving Test (RP 30) is an instrument designed to assess cognitive speed and flexibility [31]. It consists of 30 items that require the mobilisation of multiple cognitive abilities. Each item requires interpreting logical relationships within a graphical structure to determine whether specific conditions are met, with difficulty increasing progressively. The task engages core components of intelligence, including reasoning, spatial ability, numerical ability, and memory. The total score obtained in the Portuguese adaptation of the RP-30 ranges from -79 to 71, demonstrating high discriminative capacity and suitability for detecting variations in cognitive performance, particularly those associated with fatigue [31]. Regarding psychometric properties, the internal consistency of the instrument ranges between .91 and .95, indicating a very high level of reliability. These values suggest strong homogeneity among items and high precision in measuring the construct under assessment [32].

2.3 Loaded March Protocol

The activity of walking with load was carried out on May 6, 2024, at the Portuguese Military Academy - Amadora Campus, with the aim of inducing physical fatigue in the participants, constituting the first moment of data collection of the present research. The course consisted of four consecutive laps of a previously defined route, with two kilometres (km) per lap, totalling 8 km. A positive accumulated difference in altitude of approximately 125 m was estimated along the route. The lap structure allowed a more effective control of the participants during the exercise, ensuring the uniformity of the procedure. Throughout the march, the participants carried a total load of 30 kg, consisting of a 20 kg backpack (fully loaded), an FN SCAR-L automatic rifle (3,4 kg) and other individual equipment, namely uniform, boots, helmet, combat belt, tactical vest and hydration system. The hydration system consisting of a flexible reservoir (2 l) carried on the back and connected to a tube with a valve, allowing water to be ingested while walking without the need to stop. The uniformity of the load carried by the participants was guaranteed by prior weighing using a suspension scale, an instrument suitable for measuring the total weight of the individual equipment. This procedure allowed to

ensure the standardization of the conditions of physical exertion imposed during gait, contributing to the internal validity of the study.

The march took place under a self-imposed maximum effort regime, with the participants being given the guidance to complete the course at the highest speed they could maintain. This approach allowed adjusting the effort to the individual physical capacity, inducing different levels of fatigue, in coherence with the objectives of the study. After the last lap, the participants immediately moved to a room where the RP-30 cognitive test was performed. This final displacement was integrated into the protocol in order to ensure that the cognitive test was applied under the direct effect of accumulated physical fatigue.

In order to ensure safety and control of the procedure as well as reliability of monitoring, participants were organised into four groups. HR was continuously monitored using a Polar H10 chest strap, connected to the Polar Flow app, both developed by Polar Electro Oy (Kempele, Finland), using each participant's personal smartphones.

2.4 RP-30 Test Protocol

The RP-30 test was administered at two distinct time points: the first immediately after a loaded march, under fatigue conditions, and the second under resting conditions. Both administrations took place in a classroom setting, ensuring consistency of environmental conditions. Under fatigue conditions, as participants completed the march at different times, the test was administered individually, immediately upon each soldier's arrival. The execution time was recorded separately for each participant, ensuring procedural standardisation. Under resting conditions, defined by the absence of relevant physical activity on the same day, the test was administered collectively, with a single start and end time for all participants. In both conditions, the instrument was administered in paper-and-pencil format under the supervision of psychologists from the Psychopedagogical Support Section of the Military Academy.

2.5 Physiological Load Indicators

2.5.1 Heart Rate Reserve

With regard to internal load, heart rate (HR) is widely recognized as the method of choice for its quantification [33], allowing a continuous assessment of the intensity of physical effort and the physiological response of the body throughout the different phases of the activity. Derived from this parameter, the heart rate reserve (HRR) is a relevant measure for the assessment of internal load, showing significant positive relationships with variables associated with the volume and intensity of the effort performed, such as the distance travelled, reliably reflecting the physiological response to effort. HRR results from the difference between maximum heart rate (HR_{max}) and resting heart rate (RHR), serving as a basis for calculating %HRR [34], used as an indicator of the relative intensity of exertion as a function of individual ability [35].

2.5.2 Edwards Training Impulse (TRIMP) – Summated Heart Rate

The method we used to quantify internal training load is the summated heart rate zone (sumHR) model, proposed by Edwards (1993) [36]. This model, developed from the TRIMP concept, calculates the physiological load based on the time spent in different HR intensity zones. One of the main advantages of Summated Heart Rate (sumHR) is its ability to proportionally reflect the increase in physiological stress associated with higher intensities, by assigning progressively higher weights to higher HR zones. This characteristic allows a more accurate assessment of the physiological impact of exercise. In a military context, sumHR proves to be a valid and effective tool for monitoring the training load, allowing a practical and rigorous assessment of the physiological demand inherent to military activities [37].

The sumHR model, also known as Edwards' TRIMP, described in Table 1, defines five intensity zones based on HRmax percentages, associating each of them with an increasing level of physiological intensity and a proportional weighting factor [38].

Table 1. Intensity zones of the Edwards' TRIMP (sumHR)

$\%HR_{max}$	Weighting Factor	Exercise Intensity
50 – 59%	1	Very Low
60 – 69%	2	Low
70 – 79%	3	Moderate
80 – 89%	4	Hard
$\geq$ 90%	5	Very Hard

2.6 Decision Making Indicators

The objective of this study is to contribute to addressing an existing gap in the Portuguese Army regarding the relationship between physical fatigue and cognitive performance in an operational environment. Through the implementation of experimental protocols and the articulation between objective physiological indicators and cognitive performance data, it is possible to assess, in a controlled manner, the effects of fatigue on decision-making.

For this purpose, two indexes are proposed, the Decision-Making Speed Index and the Decision-Making Performance Index, which quantify, respectively, the speed and quality of decisions under fatigue providing standardized metrics for future application, which represent a methodological approach that allows direct comparison between the two experimental states, thereby enhancing the ability to identify the effects of fatigue on cognitive performance and promoting the adaptation of decisions to physical and mental states.

2.6.1 Decision-Making Speed Index

The ratio between the median value of responses obtained (MRVO) and the maximum number of possible responses (30) is considered to define the Decision-Making Speed Index (DMSI), computed under the fatigued condition and in the resting condition by:

$$DMSI = \frac{MRVO}{30} \times 100\%$$

This DMSI index allow to measure the impact, in percentage points, on military performance in terms of decision-making speed when under fatigue.

2.6.2 Decision-Making Performance Index

The ratio between the mean score obtained (MSO) in the RP-30 test and the maximum score that can be obtained in the test (71), is considered to define the Decision-Making Performance Index (DMPI), computed under the fatigued condition and in the resting condition by:

$$DMPI = \frac{MSO}{71} \times 100\%$$

This DMPI index provides an indication of the performance attained as a percentage of the maximum possible score in the test.

2.7 Statistical Methodology

A sample of 4th year students from the Portuguese Military Academy was selected from the Military Academy cadets at the Military Academy, who are physically and cognitively prepared to perform demanding tasks. It consists of 36 cadets who voluntarily participated in this study, composed exclusively of male military personnel, as a methodological approach designed to ensure the homogeneity and representativeness of the sample. The chosen group was selected for its suitability to the research objectives, as it provides the conditions required to analyse the influence of fatigue on decision-making in a military context.

The protocol implementation was carried out at two different moments, in May 2024, the RP-test after an 8km loaded march (with fatigue) and in June 2024, the RP-30 test (without fatigue), in order to safeguard the independence of the measurements and, consequently, the internal validity of the results.

This is a quantitative study as it uses information collected in the field, based on a previously defined sample of soldiers, considering standardised information collection instruments as sources of analysis. The data collected were structured and statistically analysed using IBM SPSS Statistics 30 and Microsoft Excel software, exploring the relevant variables [39]. The data was analysed using various descriptive statistical methods, combined with adequate statistical inference techniques. In order to compare the results of the RP-30 test in two conditions (with and without fatigue) in two aspects, speed and quality of decision-making, parametric hypothesis tests were used to analyse differences in the mean population parameter by implementing Student's t-test [39],

with a significance level of 10%. The assumption of normally distribution of the data was verified using the Kolmogorov–Smirnov test. Levene's test was also considered to evaluate homogeneity of variances [40]. If the necessary assumptions were not verified, non-parametric statistical methods for hypothesis testing (median population parameter) were also implemented to compare the results in the two conditions under analysis, considering the Mann-Whitney test [39].

3 Results Analysis and Discussion

The participants have been chosen among the Military Academy students who have been considered suitability to the research objectives and constitute a representative sample. A sample of 36 students of 4th year from the Portuguese Military Academy was selected. The participants were between 21 and 30 years old with an average of 23 years and a standard deviation (sd) of 2,2 years. Weight was between 62 kg and 95 kg, with an average of 78,14 kg and a sd of 7,63 kg; The height of the subjects ranged from 1,65 m and 1,92 m with an average of 1,78 m and a sd of 0,07 m. In terms of body mass index (BMI), lies between 18,51 kg/m^2 and 28,68 kg/m^2 with an average of 24,56 kg/m^2 and a sd of 2,02 kg/m^2 (Table 2).

Table 2. Characteristics of the participants

Statistics	Age (years)	Heigh (m)	Weight (Kg)	Body mass index
Minimum	21	1,65	62	18,51
Mean	23	1,78	78,14	24,56
Standard deviation	2,2	0,07	7,63	2,02
Median	22	1,78	77	24,52
Maximum	30	1,92	95	28,68

Each participant completed the RP 30 Problem Solving Test, which assesses cognitive abilities essential to decision making, on two different measurement moments. Firstly, the individuals performed a ruck march carrying an approximate load of 30 kg, second, on another day, they completed the test in a resting condition. The data collected on the physiological parameters in the participants, namely heart rate, continuously monitored during the march, enabling calculation of internal load sumHR Zone method, as well as changes in the %HRR.

3.1 Internal Load Analysis Using Edwards TRIMP

The Edwards sumHR model, was used to estimate the internal load, based on the physiological response to exertion. The median is 269 arbitrary units (a.u.) com values ranging from a minimum of 112 a.u. to a maximum of 318 a.u. The average load recorded was 263 a.u., with a standard deviation of 40,37. The interquartile range is between 241 a.u. (Q1) and 293 a.u. (Q3), with a lower atypical value (112 a.u.).

In addition, the distribution of time (min) by the intensity zones defined in the Edwards model was analysed, as shown in Fig. 1. It was found that 50,6% of the total march time was spent in zone 4 (1172 min), while 31,95% of the time was spent in zone 5 (740 min). Together, these two zones, associated with high intensities, represented approximately 82,58% of the total time recorded. Zone 3, associated with moderate intensity, corresponded to 12,82% of the total gait time (297 min). Zones 1 and 2, associated with low intensities, together accounted for 4,62% of the total accumulated time (28 min in zone 1 and 79 min in zone 2).

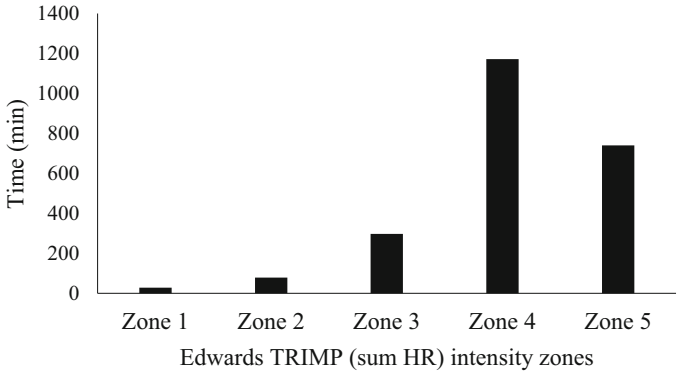


Fig. 1. Intensity zones distribution during the loaded march

3.2 %HRR Monitoring During the Loaded Military March

The distribution of %HRR over the 72 min of the loaded march is represented in Fig. 2, using box-plots each one corresponding to each min of the activity. This time interval was defined based on the time spent by the participant who took the longest to complete the course, thus allowing a complete analysis of the entire sample. Each boxplot illustrates the interindividual variation of %HRR in each min, making it possible to follow the evolution of the intensity throughout the exercise.

In general, there is a tendency for %HRR to stabilize at high values during most of the walk, reflecting the maintenance of a consistent exertion. However, the data show considerable variability among participants, with fluctuations that reach maximums above 100% and minimums below 50%, indicating moments of greater and lesser intensity. It is important to note that, at the end of the graph, there is a progressive decrease in %HRR, which does not represent the entire sample, since these datapoints correspond only to participants who were still on the march. This reduction may therefore convey an incorrect impression of a decrease in intensity in the final phase, when in reality it reflects the lower representativeness of data at these moments.

The overall median %HRR was 82,88%, with a minimum value of 43,60% and a maximum of 103,42%, which shows a predominantly high cardiovascular intensity, but with significant oscillations between individuals.

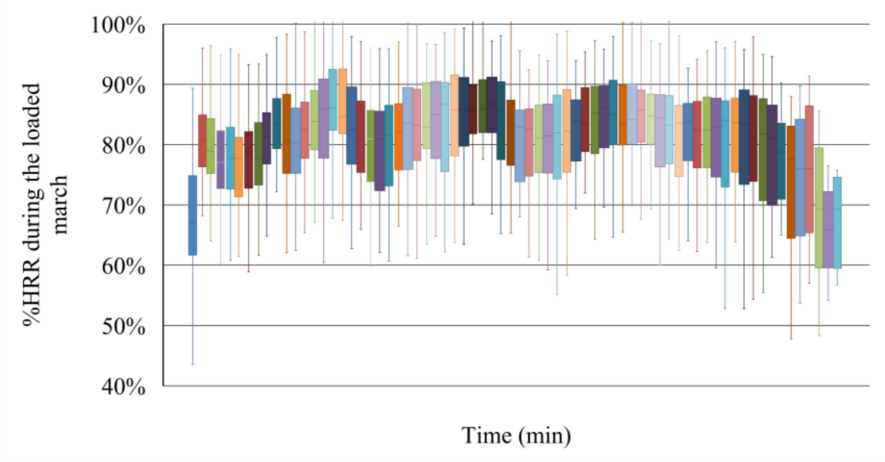


Fig. 2. Distribution of %HRR during the loaded march

3.3 Decision-Making Indicators

This study objectively reflects comparisons between the proposed experimental conditions, by means of the Decision-Making Speed Index and the Decision-Making Performance Index, that are implemented and analysed constituting a useful tool that consider the RP-30 test to quantify the impact of fatigue on decision-making.

3.4 Decision-Making Speed with and Without Fatigue

In the speed analysis, the variables number of responses given without fatigue (X_1) and number of responses given with fatigue (X_2) were considered and are represented in Fig. 3. It was observed that, under fatigue condition, participants demonstrated a reduction in the median number of responses provided during the RP-30 test. All participants completed the test within the same time frame (17 min), indicating that this difference reflects a decline in performance under fatigue. Under non-fatigued conditions, the median number of responses was 27, with a minimum value of 13 and a maximum value of 30, with a mean of 25,69 and a standard deviation of 4,65 with 50% of participants providing between 23 and 30 responses. On the other hand, after undergoing the load march, the median number of responses obtained was 20, with values ranging from 11 to 30, a mean value of 20, with a standard deviation of 5,33. In this case, 50% of participants providing between 17 and 23 responses on the RP-30 test, gradually lower in the fatigue condition when comparing the two conditions.

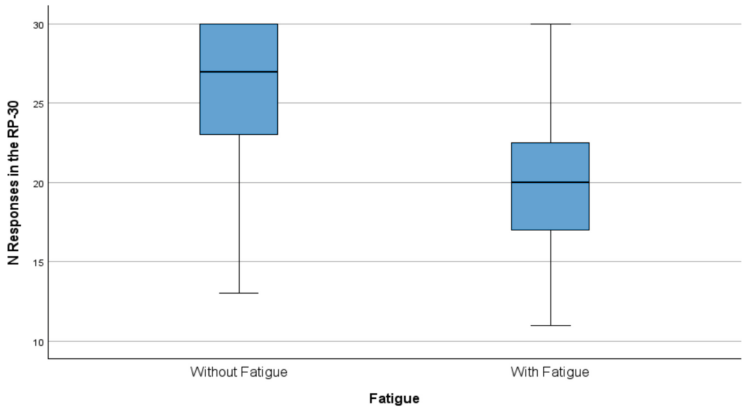


Fig. 3. Boxplot relative to N. of Responses in the RP-30 Test with and without Fatigue

According to the study objective, and taking into account that the assumption of normality for the distribution of the variables under study has not been satisfied is considered the null hypothesis, H0: The medians are the same across the proposed experimental conditions, under fatigue condition and non-fatigued condition are compared; this allows an evaluation of the fatigue and non-fatigue conditions in terms of the number of responses provided during the RP-30 test. Resorting to Mann-Whitney test, the null is rejected ($U = 278{,}5$, p-value $< 0{,}001$), therefore significant different results for the RP-30 test are provided depending on the experimental condition.

Consequently, the value obtained for the Decision-Making Speed Index was 90% under non-fatigued condition and 66,67%, under fatigue conditions, the DMSI was 66,67%. This index revealed that, when fatigued, participants experienced a 23,33% reduction in decision-making speed.

3.5 Decision-Making Quality with and Without Fatigue

Regarding decision quality, taking into account the Direct Score (DS) in the RP-30 test, the variables DS without fatigue (Y_1) and DS with fatigue (Y_2) were analysed and are represented in Fig. 4. Under non-fatigued condition, the DP scored mean of 39,48, with a standard deviation of 7,09, with a sample range of 25,74 (minimum of 24,96 and a maximum of 50,7). The median value was 39,39, obtain the values of 45,04 and 32,95 for the third and first quartiles, respectively. When experiencing fatigue, the mean score decreases by 10,71 points, reaching a value of 28,77, with a standard deviation of 8,50.

50% of participants scored between 33,345 and 23,595 in the direct scores of the test, with a median value of 39,39. Overall, it can be observed that, under fatigue conditions, there was a decrease in the DP achieved in the RP-30 test.

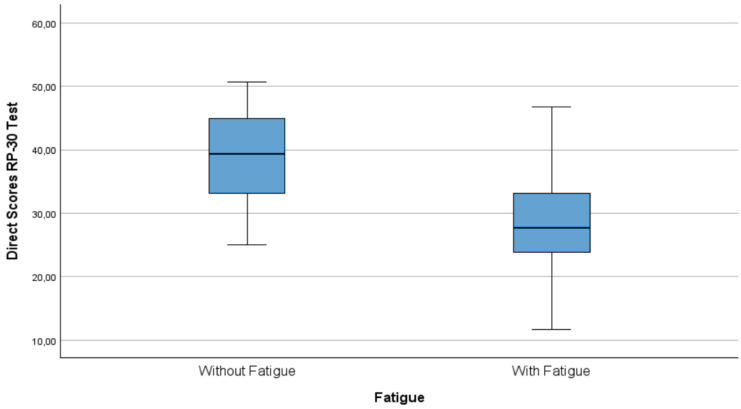


Fig. 4. Boxplot relative to Direct Scores on the RP-30 Test with and without Fatigue

From the perspective of the DP scored variables, to compare PD under non-fatigued and fatigued conditions, the differences in the means of the variables were analysed, by applying Student's t-test. The required assumptions were verified. The Kolmogorov-Smirnov test indicated that the variables Y_1 and Y_2 can be considered to originate from a normal distribution ($p > 0{,}200$). Furthermore, using Levene's test, the homogeneity of variances for variables Y_1 and Y_2 was established ($p = 0{,}532$). The application of Student's t-test ($t = 5{,}802$; $p < 0{,}001$) led to the rejection of the null hypotheses of equal expected values, concluding that the mean DP score is significantly lower under fatigue than in the non-fatigued condition.

The Decision-Making Quality Index, which reflects the performance achieved as a percentage of the maximum possible score in the test, in the non-fatigued condition is 55,61%, after the march, this value declines to 40,52%, representing a difference of 15,09 percentage points between the two conditions. The DMQI index revealed that, when fatigued, participants experienced a 15,09% reduction in decision-making quality.

4 Conclusions and Final Remarks

Decision-making in a military environment is a central process for the success of operations, requiring speed, precision, and the ability to adapt to scenarios marked by uncertainty, time pressure, and high complexity. In these contexts, military personnel are often exposed to intense physical demands, such as prolonged loaded marches, which can lead to states of physical and cognitive fatigue. This functional state, resulting from the combination of these forms of fatigue, is characterized by a decrease in the availability of mental resources, compromising the effectiveness of the processes of perception, attention, memory and reasoning as essential pillars for effective decisions in critical situations.

These results demonstrate that fatigue compromised the military's ability to maintain structured reasoning, analyse alternatives, and respond accurately, directly interfering with higher cognitive functions associated with decision-making.

The results obtained in this study contributed to fill an existing gap in the context of the Portuguese Army with regard to the relationship between physical fatigue and cognitive performance in an operational context. The construction of the DMSI and DMPI indexes represented a methodological approach that enabled a direct comparison between the two experimental states. These instruments may be useful in future studies that aim to integrate physiological and cognitive variables, especially in applied contexts. Based on the data obtained, measures with potential for practical application are also identified. In this sense, the importance of preparing service members to make decisions in conditions of fatigue is highlighted, through exercises that combine intense physical effort with situations that require information evaluation, choice between alternatives, or decision-making under pressure. More than just training resistance to effort, it is important to develop the ability to recognize the effects of fatigue on cognitive performance, promoting the adaptation of decisions to altered physical and mental states. This approach reinforces the functional integration between physical and cognitive capacities, contributing to a more effective performance in demanding operational contexts.

Notwithstanding the contributions of the present study, it is important to recognize its limitations. The sample consisted exclusively of male military personnel, in a methodological decision aimed at ensuring physiological homogeneity and controlling variables influenced by gender, such as differences in body composition, response to physical exertion and exercise tolerance.

For future research, it is recommended to broaden the sample to female military personnel, in order to allow a comparative analysis between genders and to assess possible differences in the response to fatigue. It is also suggested to increase the sample size, in order to strengthen the statistical power of the analyses and allow more refined segmentations. Finally, it is proposed to explore the relationship between different levels of physiological fatigue and performance in decision-making, with a view to deepening the knowledge about the gradual effects of fatigue on critical cognitive functions in an operational environment.

Acknowledgments. This work was partially funded by national funds through the Military Academy Research Centre (CINAMIL) and the Military Readiness Lab. This work was partially supported by national funds through the FCT – Fundação para a Ciência e Tecnologia, I.P., under the scope of the projects UID/00297/2025 (https://doi.org/10.54499/UID/00297/2025) and UID/PRR/00297/2025 (https://doi.org/10.54499/UID/PRR/00297/2025) (Centre for Mathematics and Applications).

Disclosure of Interests. The authors have no conflicts of interest to declare**.**

References

1. De Boisboissel, G.: New technologies and decision-making for the military. In: Advances in Decision Making (2021)
2. Deveci, M., Kuvvetli, Y., Akyurt, İ.Z.: Survey on military operations of fuzzy set theory and its applications (2020)
3. Vaes, A.W., et al.: Physical and mental fatigue in people with non-communicable chronic diseases. Ann. Med. **54**, 2521–2533 (2022)

4. Alkhawaldeh, I.: The effect of mental fatigue on the accuracy of the direct free kick in terms of some kinematic variables for football players. Asian J. Sports Med. (2023)
5. Nasirzadeh, F., Mir, M., Hussain, S., Sustainability, M., Darbandy, T.: Physical fatigue detection using entropy analysis of heart rate signals. Sustainability 2020•mdpi.com (2020)
6. Torri, F., Lopriore, P., Montano, V., Siciliano, G.: Pathophysiology and management of fatigue in neuromuscular diseases. Int. J. Molecular Sci. 2023•mdpi.com (2023)
7. Jin, H., et al.: Relationship between physical fatigue and mental fatigue based on multimodal meas-urement under different load levels. Ergonomics 2024•Taylor & Francis **67**, 1748–1763 (2024). https://doi.org/10.1080/00140139.2024.2364667
8. Nindl, B., Leone, C., Tharion, M.: Physical performance responses during 72 h of mili-tary operational stress. Med. Sci. Sports & Exercise, 2002•ovid.com (2022)
9. Luxton, D., Greenburg, D., Ryan, J., Niven, A., Wheeler G, Mysliwiek V: Prevalence and impact of short sleep duration in redeployed OIF soldiers. Sleep, 2011•academic.oup.com (2011)
10. Sekel, N.M., et al.: Mili-tary tactical adaptive decision making during simulated military operational stress is in-fluenced by personality, resilience, aerobic fitness, and neurocognitive function. Front. Psychol. **14** (2023). https://doi.org/10.3389/FPSYG.2023.1102425/FULL
11. Johnson, D.C., et al.: Modifying resilience mechanisms in at risk individuals: a controlled study of mindfulness training in Marines preparing for deployment. Am. J. Psychiatry **171**, 844–853 (2014). https://doi.org/10.1176/APPI.AJP.2014.13040502
12. Flood, A., Keegan, R.J.: Cognitive resilience to psychological stress in military per-sonnel. Front. Psychol. **13** (2022). https://doi.org/10.3389/fpsyg.2022.809003
13. Weeks, S., McAuliffe, C., DuRussel, D., Pasquina, P.: Physiological and psychological fatigue in extreme conditions: the military example. Elsevier (2010)
14. Bustos, D., Guedes, J., Vaz, M., Pombo E: Non-invasive physiological monitoring for physical exertion and fatigue assessment in military personnel: a systematic review. Int. J. Environ. Res. Public Health 2021•mdpi.com (2021)
15. Oeschger, R., Roos, L., Wyss, T., Buller, M.: Influence of soldiers' cardiorespiratory fitness on physiological responses and dropouts during a loaded long-distance march. Military Med. 2023•academic.oup.com. (2023)
16. Walsh, D., Palevo, G., Polascik, M., Slaton, J.: Physiological differences of US army cadets during a loaded and unloaded 6-mile ruck march. researchgate.net (2020)
17. Weyden, M. van der, Merrigan, J.J., Newman, K., Hahn, J., Martin, J.: Army Combat Fitness Test Scores moderate cognitive function improvements after a ruck march: a hi-erarchical linear model approach. J. Strength & Conditioning Res. 2022•journals.lww.com (2022)
18. Gheondea, D.: Military Mobility – A Present-Day Requirement –. Romanian Military Thinking. pp. 44–63 (2022)
19. Costa, K. da, Cantelon, J., Hussey, E., Ward, N.: Psychological responses to ruck marches during military training. researchgate.net (2022)
20. Bloch, A., Steckenrider, J., RA Zifchock, Freisinger, G., Bode, V., Elkin-Frankston, S.: Effect of fatigue on movement patterns during a loaded ruck march. Military Medicine, 2024•academic.oup.com (2024)
21. Giles, G.E., Hasselquist, L., Caruso, C.M., Eddy, M.D.: Load carriage and physical exertion influence cognitive control in military scenarios. Med. Sci. Sports Exerc. **51**, 2540–2546 (2019). https://doi.org/10.1249/MSS.0000000000002085
22. Bloch, A., Steckenrider, J., Zifchock RA: A systematic review on the effect of cognitive fatigue in teams. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 68, no. 1, pp. 1287–1291. Sage CA: Los Angeles, CA: SAGE Publications (2024). https://doi.org/10.1177/10711813241275515
23. Aljahni, M.: The effect of fatigue on cognitive tasks among young adults. Int. J. Phys. Therapy Res. Pract. 2024•inkwellinfinite.com. **3**, 425–435 (2024)

24. Jørgen, D., Jepsen, R., Riis Jepsen, J., Zhao, Z., Van Leeuwen, W.M.A.: Seafarer fa-tigue: a review of risk factors, consequences for seafarers' health and safety and options for mitigation. Int. Maritime Health 2015•journals.viamedica.pl. **66**, 106–117 (2015). https://doi.org/10.5603/IMH.2015.0024
25. Lindner, C., Nagy, G., Roell, L., Zitzmann, S.: Investigating the impact of perceived mental fatigue on sustained attention performance: a latent growth curve analysis taking social desirability into account. Cogn. Emot. **39**, 1976–1987 (2025). https://doi.org/10.1080/02699931.2025.2468281
26. Hepler, T.: Decision-making in sport under mental and physical stress. academia.edu. (2015)
27. Jia, H., Lin, C., Wang EM: Effects of mental fatigue on risk preference and feedback processing in risk decision-making. Sci. Rep. 2022•nature.com **12**, 10695 (2022). https://doi.org/10.1038/s41598-022-14682-0
28. Hockey, R.: The psychology of fatigue: Work, effort and control (2013)
29. Hopstaken, J., Linden, D., Van Der, Bakker, A.B., Kompier, M.A.: A multifaceted investigation of the link between mental fatigue and task disengagement. Psychophysiology, 2015•Wiley Online Library, vol. 52, pp. 305–315 (2015). https://doi.org/10.1111/psyp.12339
30. Linden, D. Van der, Frese, M., Meijman, T.F.: Mental fatigue and the control of cognitive processes: effects on perseveration and planning. Elsevier (2003)
31. Seisdedos, N.: RP-30: Resolução de Problemas–Manual técnico. Editora Hogrefe. (2020)
32. Barros, E., Kausel, E.E., Cuadra, F., Díaz, D.A.: Using general mental ability and personality traits to predict job performance in three Chilean organizations. Int. J. Sel. Assess. **22**, 432–438 (2014)
33. Espasa-Labrador, J., Fort-Vanmeerhaeghe, A., Montalvo, A., Carrasco-Marginet, M., Irurtia, A., Calleja-Gonzalez, J.: Monitoring internal load in women's basketball via sub-jective and device-based methods: a systematic review. mdpi.com (2023)
34. Sobolewski, E.: The relationships between internal and external load measures for divi-sion I college football practice. mdpi.com2020•mdpi.com (2020)
35. Swain, D., Leutholtz, B., King, M., Haas, L.: Relationship between% heart rate reserve and% VO2reserve in treadmill exercise. Med. Sci. Sports & Exercise 1998•ovid.com (1998)
36. Edwards, S.: High performance training and racing. In: The Heart Rate Monitor Book (1993)
37. Canino, M., et al.: Quantifying training load during physically demanding tasks in US Army soldiers: a comparison of physiological and psychological measurements. Military Med. 2020•academic.oup.com. (2020)
38. Mckee, J.R., De Marco, K., Girard, O., Peiffer, J.J., Scott, B.R.: Effects of blood flow restriction on internal and external training load metrics during acute and chronic short-term repeated-sprint training in team-sport athletes. J. Sports Sci. **43** (2025). https://doi.org/10.1080/02640414.2025.2457863
39. Casella, G., Berger, R.L.: Statistical Inference, Second Edition. Statistical Inference, Second Edition., pp. 1–535 (2024). https://doi.org/10.1201/9781003456285
40. Yonar, A., Yonar, H., Demirsoz, M., Tekindal, M.: A comparative analysis for homo-geneityof variance tests. J. Sci. Arts. **24** (2023). https://doi.org/10.46939/j.sci.arts-24.2-a06



A Framework for the Statistical Analysis of Fixed Effects Models Using *R*, Illustrated with an Example in Biometry

Teresa A. Oliveira[1,2] and Maria Filomena Teodoro[3,4(✉)]

[1] CEAUL, Faculdade de Ciencias, Universidade de Lisboa, 1649-014 Lisbon, Portugal
[2] Departamento de Ciencias e Tecnologia, Universidade Aberta, 1269-001 Lisbon, Portugal
[3] CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, 2810-001 Almada, Portugal
mteodoro64@gmail.com
[4] CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Tecnico, Lisbon University, 1048-001 Lisbon, Portugal

Abstract. To ensure meaningful insights experiments must be thoughtfully designed to support structured comparisons across treatments. A key tool in this context is the one-way ANOVA, which evaluates whether observed differences in group means reflect true treatment effects rather than random variation. Exploring the importance of One-Way ANOVA in biometrical studies, discussing its assumptions, applications and advantages, a structured and accessible framework is presented for the statistical analysis of fixed effects models using R, illustrated with an application in biometry.

Designed to support both learners and practitioners, the framework combines clear pedagogical guidance with methodological rigor. It introduces key concepts such as model specification, estimation, and inference, while emphasizing reproducibility and interpretability through annotated ***R*** code considering a practical example. Beyond its educational value, the framework contributes a practical and computationally efficient approach to hypothesis testing, model validation, and post-hoc analysis. By bridging statistical theory with applied research, a valuable resource is here provided for students, educators, and biometricians seeking to conduct robust and transparent fixed effects analyses.

1 Introduction

The ability to draw valid and reproducible conclusions depends fundamentally on the quality of the experimental design [1–4]. When comparing multiple treatments or conditions, rather than performing multiple pairwise comparisons, which would be inefficient and statistically unsound, the One-Way ANOVA offers a more rigorous and reliable solution. Structured experimental frameworks are essential to ensure that observed differences are attributable to the factors under study, rather than to random variation. Beyond one-way analysis of variance (ANOVA), there is a fixed effects model that, under controlled conditions, allow the researchers to assess whether the means of different groups differ significantly [1, 5]. The one-way layout ANOVA is widely applied in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 132–146, 2027.
https://doi.org/10.1007/978-3-032-30521-3_9

biological research once it is particularly suited for experiments where a single categorical factor (e.g., treatment type, genotype, or environmental condition) is used to classify observations into distinct groups. Under the assumption of fixed treatment effects, this model partitions the total variability in the data into components attributable to between-group and within-group variation, enabling formal hypothesis testing regarding group differences [6, 7].

Despite its conceptual simplicity, the correct application and interpretation of fixed effects models require careful attention to model assumptions, estimation procedures, and post-hoc analysis. A structured and reproducible framework for the statistical analysis of fixed effects models using the R programming language, including an illustrative example, is presented. The goal is to provide both a pedagogical resource for learners and a practical toolkit for researchers conducting experimental studies in the life sciences [8–14].

2 Methodological Foundations of ANOVA One-Way Layout: The Fixed Effects Model

The One-Way ANOVA methodology [1, 6–8], originally developed by Sir Ronald A. Fisher in the early 20th century, revolutionized experimental design and statistical inference. The resulting F-statistic follows the F-distribution, named in Fisher's honor, and serves as the basis for drawing conclusions about the presence or absence of significant differences among group means.

One-Way ANOVA is based on several key assumptions: normality, homoscedasticity (equal variances), and independence of observations. These assumptions ensure the validity of the F-test, which is used to determine whether the observed differences among group means are statistically significant. The method involves partitioning the total variability in the data into two components: between-group variability and within-group variability. The F-ratio, calculated as the ratio of between-group variance to within-group variance, serves as the test statistic. A significantly large F-value suggests that the group means are not all equal.

In experiments with one factor the data or observations consist of k independent groups of samples (levels) and in each group n_i measurements are taken.

The response observed for each treatment (or observations in the i-th sample) is a random variable Y_{ij} whose concrete values are represented by y_{ij}, with $i = 1,2,\ldots,k$; $j = 1,2,\ldots, n_i$.

If the number of observations n_i is the same, the samples have all the same dimension and that is known as the balanced case. In general model we consider the existence of samples with different dimensions, this the case designated as unbalanced.

2.1 The One-Way Fixed Effects Model and the Data Structure

The appropriate model to describe this type of data is a linear statistical model of the type:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ \mathrm{i} = 1, 2, \ldots, \mathrm{k};\ \mathrm{j} = 1, 2, \ldots n_i,$$

where:

$Y_{ij} \rightarrow$ random variable representative of j-th observation of i-th treatment;

$\mu \rightarrow$ Parameter common to all treatments and designated by global mean;

$\tau_i \rightarrow$ Single parameter for treatment i designated by the effect of i-th treatment;

$\varepsilon_{ij} \rightarrow$ Random component of the error or residuals.

The main objective of the problem is to compare the mean values $\tau_1, \tau_2, ..., \tau_k$ in order to investigate the existence of significant differences between them, and the analysis can be extended to the estimation of the effects of the treatments.

- In order to carry out the hypothesis tests, the following assumptions must be made regarding the random component:
- The errors, ε_{ij}, must be independent normal random variables, of null mean value and variance σ^2, constant for all factor levels.

The one-way ANOVA can be adapted for the investigation of two different situations regarding the effects of the treatments:

- In the first case the experimenter can initially define his interest in testing only the specific treatments in question and, obviously, draw conclusions only with respect to the treatments under study. These conclusions cannot subsequently be adapted or extrapolated to other similar treatments. The investigator can go further and obtain estimators for the parameters of the model, namely μ, τ_i e σ^2. This type of situation is known as the Fixed Effects Model.
- In a second case, one can consider that k treatments are obtained from a wide population of treatments. These situations commonly occur in more complex experiments, involving a large number of treatments to be compared, the study of the entire population involved being neither practical nor economically feasible. In this case it will be desirable to extend the conclusions drawn from the sample to all treatments of the population, whether involved in the analysis or not. The random variables are represented by τ_i and the objective is more related with testing the variability. This type of situation is known by Random Effects Model.

This paper is devoted to the framework of the first case.

The data structure typically consists of observations of a problem with o factor can be arranged in a table with k columns and n_i rows.

Typical data structure for experiments with one factor and k levels

Treatments					
Observations	**1**	**2**	**3**	**…**	**k**
	y_{11}	y_{12}	y_{13}	⋯	y_{1k}
	y_{21}	y_{22}	y_{23}	⋯	y_{2k}
	⋮	⋮	⋮	⋮	⋮
	$y_{n_1 1}$	$y_{n_2 2}$	$y_{n_3 3}$	⋯	$y_{n_k k}$

2.2 The Fundamentals of Statistical Analysis

In the fixed effects model the effects of treatments τ_i are usually defined as deviations from the global mean and it is assumed that:

$$\sum_{i=1}^{k} \tau_i = 0$$

Consider that:
y_{io} is the sum of the total of observations of ith treatment;
$y_{i\bullet}$ is the mean of observations of i-th treatment;
$y_{\bullet\bullet}$ is the global mean of observations.
Thus:

$$y_{io} = \sum_{j=1}^{n_i} y_{ij}, \, y_{i\bullet} = \frac{y_{io}}{n_i}, \; i = 1, 2, ..., k$$

$$y_{oo} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}, \, y_{\bullet\bullet} = \frac{y_{oo}}{n}, \, n = \sum_{i=1}^{k} n_i$$

The mean, or expected value, for the ith treatment is given by:

$$E(Y_{ij}) = \mu_i = \mu + \tau_i, \, i = 1, ..., k.$$

Consequently, the mean of the ith treatment consists of the sum of the global mean with the effect of the ith treatment. The hypothesis to be tested are:
$H_o : \mu_1 = \mu_2 = ... = \mu_k$, versus $H_1 : \mu_i \neq \mu_j$, for at least one pair (i, j).
Note that, under $H_o : \mu_1 = \mu_2 = ... = \mu_k$, all treatments have μ as common mean.
Another equivalent way of writing these hypotheses is:
$H_o : \tau_1 = \tau_2 = ... = \tau_k = 0$, versus $H_1 : \tau_i \neq 0$, for at least some i.
Note that, testing the equality of treatments, or testing that the effects of treatments are null, is the same. As already mentioned, the appropriate procedure for testing the equality of treatments is called analysis of variance and, the general procedures will be described below.
The designation "Analysis of variance" is due to the fact that all this analysis is based on a partition of the total variability of observations in components. The total sum of squares, represented by SQT, measures the overall variability of the data.
Thus:

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - y_{\bullet\bullet})^2$$

Note that, if SQT is divided by the number of corresponding degrees of freedom, that is, $n_1 + n_2 + ... + n_k - 1 = n - 1$, the obtained result is the estimated variance for the sample taking into account the n observations.

The total sum of squares can be partitioned as follows:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-y_{\bullet\bullet})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}[\,(y_{i\bullet}-y_{\bullet\bullet})+(y_{ij}-y_{i\bullet})\,]^2$$

($y_{i\bullet}$ is simply added and subtracted from the expression and so the result does not change).

Developing this last expression as square of sums, it comes:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}[\,(y_{i\bullet}-y_{\bullet\bullet})+(y_{ij}-y_{i\bullet})\,]^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i\bullet}-y_{\bullet\bullet})^2$$
$$+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i\bullet}-y_{\bullet\bullet})(y_{ij}-y_{i\bullet})+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-y_{i\bullet})^2$$

Note that:

$$2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i\bullet}-y_{\bullet\bullet})(y_{ij}-y_{i\bullet})=0.$$

And, moreover,

$$\sum_{j=1}^{n_i}(y_{ij}-y_{i\bullet})=y_{io}-n_iy_{i\bullet}=y_{io}-n_i\frac{y_{io}}{n_i}=y_{io}-y_{io}=0$$

The total variability can be written as:

$$\sum_{\mathrm{i}=1}^{\mathrm{k}}\sum_{\mathrm{j}=1}^{\mathrm{n}_i}(\mathrm{y}_{\mathrm{ij}}-\mathrm{y}_{\bullet\bullet})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i\bullet}-y_{\bullet\bullet})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-y_{i\bullet})^2$$

Resuming, the total variability can be expressed as a partition of:

(i) The sum of squares of differences between the means of the samples (or treatments) and the global mean;

and

(ii) The sum of squares of differences between observations within the samples (or treatments) and average treatments.

Thus, the difference between the means of treatments observed and the overall mean is a measure of the differences between treatment means.

Symbolically, the above expression can be written as SQT = SQA + SQE, where:

SQT – denotes the total sum of squares

SQA – denotes the sum of squares due to the treatments

SQE – denotes the sum of squares of errors or residuals

The data consist of n observations in total, where SQT has n – 1 degrees of freedom; there are k levels of the factor (k different treatments or samples to compare) and thus

SQA has k − 1 degrees of freedom. As for the error, it is known that for each treatment there is $n_i - 1$ degrees of freedom. Since there are k treatments, the number of degrees of freedom for the error is given by:

$$(n_1 - 1) + (n_2 - 1) + ... + (n_k - 1) = \sum_{i=1}^{k} n_i + \sum_{i=1}^{k} 1 = n - k.$$

The basic identity of the analysis of variance consists of two estimates of variance, one based on the variability within the treatments and another variability between treatments.

If there are no differences between the means of the treatments, these two estimates should be very similar and if they are not, this is a reason to suspect that the observed differences will be caused by differences in the means of the treatments. Consider the following quantities:

$$QMA = \frac{SQA}{k - 1} \quad \text{and} \quad QME = \frac{SQE}{n - k},$$

so that QMA represents the samples mean square and QME represents the error mean square. The hypothesis test for differences between treatment means can be established by comparing the variance estimators QMA and QME. The ratio between QMA and QME is therefore called the variance ratio.

As for the errors, the normality, independence, homoscedasticity and expected null value were assumed, for the variables representative of the observations, Y_{ij}, it is assumed that they are identically distributed, mutually independent, and follow approximately the normal distribution, with mean $\mu + \tau_i$, and have the same variance σ^2 [9–12].

Once that the sum of squares for the total is a sum of squares of normal random variables, then,

$\frac{SQT}{\sigma^2} \approx \chi^2_{n-1}$, $\frac{SQE}{\sigma^2} \approx \chi^2_{n-k}$, $\frac{SQA}{\sigma^2} \approx \chi^2_{k-1}$ (considering H_0 is true).

However, once that SQT = SQA + SQE, then the three sums of squares are not independent. It is necessary to resort to a Theorem, Cochran's Theorem, to establish the independence of SQA and SQE. Thus, if H_0 is true, F_0 follows the F distribution with k – 1, n – k degrees of freedom and is defined by:

$$F_0 = \frac{SQA/k - 1}{SQE/n - k} = \frac{QMA}{QME}$$

F_0 is the statistic to test the hypothesis that there are no differences between the means of the treatments. The decision rule is to reject H0 at the significance level α, if $F_0 > F_{k-1,n-k}(\alpha\%)$. The correspondent ANOVA table to test one factor for fixed effects models is:

ANOVA one-way, fixed effects model: general case

O.V	G.L	S.Q	Q.M	R.V
Between samples (treatments)	k – 1	$SQA = \sum_{i=1}^{k} n_i y_{i\bullet}^2 - n y_{\bullet\bullet}^2$	QMA	$F_0 = \frac{QMA}{QME}$

(*continued*)

(*continued*)

O.V	G.L	S.Q	Q.M	R.V
Residual	n – k	SQE = SQT-SQA	QME	
Total	n – 1	$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - ny_{\bullet\bullet}^2$		

In computational terms it is simpler to use the expressions presented in the previous table for the calculation of SQT, SQA and SQE, respectively. These expressions are easily deduced from the corresponding equivalent expressions given above. In the balanced case, where the samples have equal size, the data can be arranged analogously in a table with r lines (fixed number of observations per treatment or sample) and k columns (number of treatments or samples) and the computations can be simplified as follows:

Observations in experiments with one factor – balanced case

Treatments					
Observations	**1**	**2**	**3**	**…**	**K**
	y_{11}	y_{12}	y_{13}	⋯	y_{1k}
	y_{21}	y_{22}	y_{23}	⋯	y_{2k}
	⋮	⋮	⋮	⋮	⋮
	y_{r1}	y_{r2}	y_{r3}	⋯	y_{rk}

$$y_{io} = \sum_{j=1}^{r} y_{ij} \text{ and } y_{i\bullet} = \frac{1}{r}\sum_{j=1}^{r} y_{ij}$$

$$SQA = r\sum_{i=1}^{k} y_{i\bullet}^2 - ny_{\bullet\bullet}^2 \text{ with } SQT = \sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - y_{\bullet\bullet})^2, \text{ and } r\ \ k\ =\ n.$$

Thus, the ANOVA table can be simplified:

ANOVA One-way, fixed effects model: balanced case

O.V	G.L	S.Q	Q.M	R.V
Between samples (treatments)	k – 1	$SQA = r\sum_{i=1}^{k} y_{i\bullet}^2 - ny_{\bullet\bullet}^2$	QMA	$F_0 = \frac{QMA}{QME}$
Residual	n – k	SQE = SQT-SQA	QME	
Total	n – 1	SQT		

2.3 Statistical Inference for Error

While QME is used to estimate σ^2, QMA is also an estimator of σ^2 under the null hypothesis, that there are no differences between treatments (H0:$\tau_i = 0$).

Consider the expected values for the mean squares of the error:

$$E(QME) = E\left(\frac{SQE}{n-k}\right) = \frac{1}{n-k}E\left[\sum_{i=1}^{k}\sum_{j=1}^{r}(y_{ij} - y_{i\bullet})^2\right] = \frac{1}{n-k}E\left[\sum_{i=1}^{k}\sum_{j=1}^{r}y_{ij}^2 - \frac{1}{r}\sum_{i=1}^{k}y_{io}^2\right].$$

Note that $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ and that, $y_{io} = \sum_{j=1}^{r}\mu + \tau_i + \varepsilon_{ij}$, and so:

$$E(QME) = \frac{1}{n-k}E\left[\sum_{i=1}^{k}\sum_{j=1}^{r}(\mu + \tau_i + \varepsilon_{ij})^2 - \frac{1}{r}\sum_{i=1}^{k}\left(\sum_{j=1}^{r}\mu + \tau_i + \varepsilon_{ij}\right)^2\right]$$

After some lengthy calculations, developing the squares and taking the expected values of the terms in parentheses, and replacing the terms of the square of the error by σ^2, it finally turns out to be:

$$E(QME) = \frac{1}{n-k}\left[n\mu^2 + r\sum_{i=1}^{k}\tau_i^2 + n\sigma^2 - n\mu^2 - r\sum_{i=1}^{k}\tau_i^2 - k\sigma^2\right] = \sigma^2.$$

Similarly it can be proved that $E(QMA) = \sigma^2 + \frac{r\sum_{i=1}^{k}\tau_i^2}{k-1}$.

Finally, in cases for which the analysis of variance test is significant, multiple comparisons can be performed in order to analyze such differences.

2.4 The Scheffé Multiple Comparisons Test

For the analysis of the multiple comparisons, the statistic of Scheffé test is given by:

$T_S = \frac{\left|\hat{\tau}_i - \hat{\tau}_{i'}\right|}{\sqrt{\frac{SQE}{n-k}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}}$, which is equivalent to use: $T_S = \frac{\left|\hat{\tau}_i - \hat{\tau}_{i'}\right|}{\sqrt{QME\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}}$,

where $T_S \approx \sqrt{(k-1)F_{k-1,n-k}(\alpha)}$, and $F_{k-1,n-k}(\alpha)$ represents the value of the distribution F with (k – 1, n – k) degrees of freedom and significance level α.

The null hypothesis can be written as:

$H_0 : \tau_i = \tau_{i'}$,

and each comparison case is rejected if $T_S \geq \sqrt{(k-1)F_{k-1,n-k}(\alpha)}$, or:

$$|\hat{\tau}_i - \hat{\tau}_{i'}| \geq \sqrt{(k-1)F_{k-1,n-k}(\alpha)\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)\frac{SQE}{n-k}}$$

Thus, the differences between observed treatments higher than the value of the second member of the inequality are significant.

2.5 The Estimation of the Model Parameters

To estimate the general model parameters of:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ i = 1, ..., k\ ,\ j = 1, ..., n_i,$$

the Least Squares Method (LSM) is used.

Considering LSM for the estimation of μ and τ_i, it is not necessary to assume the usual assumptions of normality of the errors. The entity leading to the least squares is denoted by W and is represented by the following expression:

$$W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\varepsilon_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \mu - \tau_i)^2$$

The values that minimize W, are obtained by solving the following system of equations:

$$\begin{cases} \dfrac{\partial W}{\partial \mu} = 0 \\ \dfrac{\partial W}{\partial \tau_i} = 0, \quad i = 1, 2, ..., k \end{cases}$$

Thus,

$$\begin{cases} \dfrac{\partial W}{\partial \mu} = 0 \\ \dfrac{\partial W}{\partial \tau_i} = 0, \quad i = 1, 2, ..., k \end{cases} \Leftrightarrow \begin{cases} -2\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0 \\ -2\displaystyle\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0, \quad i = 1, 2, ..., k \end{cases}$$

And, after some simplifications, it can be deduced that:

$$\begin{cases} n\hat{\mu} + n_1\hat{\tau}_1 + n_2\hat{\tau}_2 + ... + n_k\hat{\tau}_k = y_{00} \\ n_1\hat{\mu} + n_1\hat{\tau}_1 \qquad\qquad\qquad\qquad = y_{10} \\ n_1\hat{\mu} \qquad + n_2\hat{\tau}_2 \qquad\qquad\qquad = y_{20} \\ \vdots \\ n_1\hat{\mu} \qquad\qquad\qquad + n_k\hat{\tau}_k = y_{k0} \end{cases}$$

These $k + 1$ equations at $k + 1$ unknowns, are known by least squares normal equations or simply by normal equations. The normal equations are not linearly independent, and the solution for $\mu, \tau_1, ..., \tau_k$ is not unique. The way to overcome this problem is to consider that $\sum_{i=1}^{k}\hat{\tau}_i = 0$, and, based on it, the solution for the normal equations is given by:

$$\hat{\mu} = y_{\bullet\bullet}$$

$\hat{\tau}_i = y_{i\bullet} - y_{\bullet\bullet}$, $i = 1, 2, ..., k$.

It may also be interesting to deduce a confidence interval to estimate the mean of the ith treatment and $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = y_{i\bullet}$, $i = 1,2,\ldots,k$, is an unbiased estimator for μ_i.

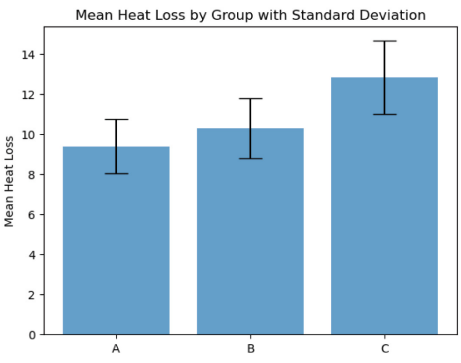
3 Illustrative Example: Classical Way and Using *R*

As an example, consider the case of a pharmaceutical laboratory conducting an experiment with the aim of comparing the heat losses of a particular type of cylinder doors, under three different temperatures and, if significant differences are confirmed, then to clarify which. Thus, these temperatures were tested in some doors, and the following data (in codified form) were obtained:

Temperatures		
A	**B**	**C**
10	12	12
11	12	12
8	11	15
10	8	10
8	10	14
	9	14
	10	

To solve this comparison problem as a first approach it is recommended to have the data visualization, and ***R*** can be used to illustrate the data graphically. Bellow, there is the representation of the ***R*** code and the resultant figure for these data. The visualization of this figure clearly allow the perception that temperature C has a higher average heat loss than the others. Is this perception significant?

```
R Console
> # Dados
> groups <- c("A", "B", "C")
> means <- c(9.40, 10.29, 12.83)
> std_devs <- c(1.34, 1.50, 1.83)
>
> # Criar o gráfico de barras com barras de erro
> barplot(
+   height = means,
+   names.arg = groups,
+   ylim = c(0, max(means + std_devs) + 1),
+   col = "skyblue",
+   main = "Mean Heat Loss by Group with Standard Deviation",
+   ylab = "Mean Heat Loss"
+ )
>
> # Adicionar barras de erro
> arrows(
+   x0 = 1:3,
+   y0 = means - std_devs,
+   x1 = 1:3,
+   y1 = means + std_devs,
+   angle = 90,
+   code = 3,
+   length = 0.1
+ )
```



The main aim is to compare the effect of the three different temperatures, A, B and C, in order to ascertain whether there are significant differences in the heat loss of a particular type of door. Assuming the usual assumptions of normality, independence and homoscedasticity, the proper model for the statistical analysis of these data is the fixed effects model, and the hypothesis, null and alternative, correspond to:

$H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 : \mu_i \neq \mu_j$, for at least one pair (i,j).

Temperatures		
A	**B**	**C**
$y_{1o} = 47$	$y_{2o} = 72$	$y_{3o} = 77$
$y_{1\bullet} = \frac{47}{5} = 9.4$	$y_{2\bullet} = \frac{72}{7} = = 10.286$	$y_{3\bullet} = \frac{77}{6} = = 12.833$

$y_{\bullet\bullet} = 10.889$; $n_1 = 5$; $n_2 = 7$; $n_3 = 6$.

$$SQA = \sum_{i=1}^{k} n_i y_{i\bullet}^2 - n y_{\bullet\bullet}^2 = 5(9.4)^2 + 7(10.286)^2 + 6(12.833)^2 - 18(10.889)^2$$

$$= 2170.528 - 18(118.570) = 36.262$$

$$SQT = \sum\sum y_{ij}^2 - n y_{\bullet\bullet}^2 = 449 + 754 + 1005 - 18(10.889) = 73.734.$$

$SQE = SQT - SQA = 73.734 - 36.262 = 37.47$

$SQE = SQT - SQA = 73.734 - 36.262 = 37.472$

ANOVA

Variation Source	Degrees of Freedom	Sum of Squares	Mean squares	Variance ratio
Between temperatures	2	36.262	18.131	7.25
Residual	15	37.472	2.498	
Total	17			

*significance level at 5%, $F_{2,15}(5\%) = 3.68$.

By observing the ANOVA table, and once that the observed value for the variance ratio is higher than the F table value, *7.25 > 3.68*, then the null hypothesis of equality between means is rejected at the *5%* significance level. At this significance level, there is statistical evidence of significant differences between temperatures in terms of heat loss.

To perform the respective ANOVA using R, the code and the results are, respectively:

```
> # Dados brutos
> A <- c(10, 11, 8, 10, 8)
> B <- c(12, 12, 11, 8, 10, 9, 10)
> C <- c(12, 12, 15, 10, 14, 14)
>
> # Criar um data frame com os dados em formato longo
> grupo <- factor(c(rep("A", length(A)), rep("B", length(B)), rep("C", length(C))))
> valor <- c(A, B, C)
> dados <- data.frame(grupo, valor)
>
> # Tabela com os valores por célula
> print("Tabela com os valores por grupo:")
```

```
[1] "Tabela com os valores por grupo:"
> print(dados)
   grupo valor
1      A    10
2      A    11
3      A     8
4      A    10
5      A     8
6      B    12
7      B    12
8      B    11
9      B     8
10     B    10
11     B     9
12     B    10
13     C    12
14     C    12
15     C    15
16     C    10
17     C    14
18     C    14
>
> # ANOVA
> anova_resultado <- aov(valor ~ grupo, data = dados)
> summary(anova_resultado)
            Df Sum Sq Mean Sq F value Pr(>F)
grupo        2  36.32  18.158   7.271 0.0062 **
Residuals   15  37.46   2.497
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
>
```

Note that, using ***R*** there is also the advantage of obtaining the correspondent *p-value,* this case, the *p-value* $= 0.0062 < 0.05$, indicates that, at the significance level of *5%*, there are significant differences between temperatures, confirming the previous classical results and conclusions.

The next step is in what concerns the parameters estimation. Thus, by the classic way, manually computing them from the model, the results are:

$$\hat{\mu} = 10.889$$

$$\hat{\tau}_1 = y_{1\bullet} - y_{\bullet\bullet} = 9.4 - 10.889 = -1.489$$

$$\hat{\tau}_2 = y_{2\bullet} - y_{\bullet\bullet} = 10.286 - 10.889 = -0.603$$

$$\hat{\tau}_3 = y_{3\bullet} - y_{\bullet\bullet} = 12.833 - 10.889 = 1.944$$

And, using ***R***, the code and the respective results are:

```
>
> # Criar os dados
> A <- c(10, 11, 8, 10, 8)
> B <- c(12, 12, 11, 8, 10, 9, 10)
> C <- c(12, 12, 15, 10, 14, 14)
>
> # Criar um vetor com os valores e outro com os grupos
> valores <- c(A, B, C)
> grupo <- factor(c(rep("A", length(A)), rep("B", length(B)), rep("C", length(C))))
>
> # Médias por grupo
> tapply(valores, grupo, mean)
       A        B        C
 9.40000 10.28571 12.83333
>
> # Efeitos estimados (diferença da média de cada grupo em relação à média global)
> media_global <- mean(valores)
> efeitos <- tapply(valores, grupo, mean) - media_global
> efeitos
         A          B          C
-1.4888889 -0.6031746  1.9444444
>
```

And, as it was expected, the previous results are confirmed by using R.

Consider to proceed with multiple comparison test, and taking into account the unbalanced data, at the significance level of *5%* the application of S (Scheffé) multiple comparison test is recommended and the respective results follow, both in classical way and using ***R***.

Comparing temperatures A and B, the results are:

- $|\hat{\tau}_1 - \hat{\tau}_2| = |-1.489 - (-0.603)| = 0.886$

$\sqrt{(3-1)\mathrm{F}_{2,15}(5\%)\left(\frac{1}{5}+\frac{1}{7}\right)(2.489)} = 2.511,$

and the conclusion is that here are no significant differences between temperatures A and B in terms of heat losses;

Comparing temperatures A and C, the results are:

- $|\hat{\tau}_1 - \hat{\tau}_3| = |-1.489 - 1.944| = 3.433$

$\sqrt{(3-1)F_{2,15}(5\%)\left(\frac{1}{5}+\frac{1}{6}\right)(2.489)} = 2.596$, and the conclusion is that there are significant differences between temperatures A and C with respect to heat losses;

Comparing temperatures B and C, the results are:

- $|\hat{\tau}_2 - \hat{\tau}_3| = |-0.603 - 1.944| = 2.547$

$\sqrt{(3-1)\mathrm{F}_{2,15}(5\%)\left(\frac{1}{7}+\frac{1}{6}\right)(2.489)} = 2.386;$

and the conclusion is that there are significant differences between temperatures B and C, with respect to heat losses. Resuming the results:

Contrasts	$\lvert\tau_i - \tau_{i'}\rvert$	Critical value (Scheffé test)	Conclusion
A versus B	0.886	2.511	A does not differ of B
A versus C	3.433	2.596	A differs of C
B versus C	2.547	2.386	B differs of C

The final answer to the Laboratory question is that the temperature C significantly differ from the temperatures A and B, which do not differ from one another, with respect to heat losses of the doors, at a significance level of *5%*.

In order to obtain the parameter estimators from an ANOVA model in ***R***, the group means can be extracted using the model.tables() function. For that, after obtaining the summary results of ANOVA in ***R***, the ***R*** code for Scheffé multiple comparisons must be added:

```
> # Estatísticas básicas
> means <- tapply(df$Temp, df$Group, mean)
> n <- table(df$Group)
> MSE <- summary(anova_model)[[1]]["Residuals", "Mean Sq"]
> df_between <- summary(anova_model)[[1]]["Group", "Df"]
> df_within <- summary(anova_model)[[1]]["Residuals", "Df"]
> alpha <- 0.05
> k <- length(means)
>
> # Comparações de Scheffé
> comparacoes <- combn(names(means), 2, simplify = FALSE)
> for (comp in comparacoes) {
+   g1 <- comp[1]
+   g2 <- comp[2]
+   diff <- abs(means[g1] - means[g2])
+   se <- sqrt(MSE * (1/n[g1] + 1/n[g2]))
+   scheffe_crit <- sqrt((k - 1) * qf(1 - alpha, df_between, df_within)) * se
+   cat(sprintf("%s vs %s:\n", g1, g2))
+   cat(sprintf("  Diferença de médias: %.2f\n", diff))
+   cat(sprintf("  Valor crítico de Scheffé: %.2f\n", scheffe_crit))
+   cat(sprintf("  Significativo? %s\n\n", ifelse(diff > scheffe_crit, "SIM", "NÃO")))
+ }
```

And the respective output is:

```
A vs B:
  Diferença de médias: 0.89
  Valor crítico de Scheffé: 2.51
  Significativo? NÃO

A vs C:
  Diferença de médias: 3.43
  Valor crítico de Scheffé: 2.60
  Significativo? SIM

B vs C:
  Diferença de médias: 2.55
  Valor crítico de Scheffé: 2.39
  Significativo? SIM
```

These results confirm those obtained previously.

An extension of the presented experimental design under the aim of biometry can be found in [13–21] for eventual interess.

References

1. Oliveira, T.A.: Estatística Aplicada. Ed. Universidade Aberta, nº287 (2004). ISBN: 972-674-444-X1
2. Branda, M.E., Kunneman, M., Meza-Contreras, A.I., et al.: Shared decision-making for patients hospitalized with acute myocardial infarction: a randomized trial. Patient Prefer. Adherence **16**, 1395–1404 (2022)
3. Eldridge-Smith, E.D., Manber, R., Tsai, S., et al.: Stepped care management of insomnia co-occurring with sleep apnea: the air study protocol. Trials **23**(1), 806 (2022)

4. Fatoye, F., Gebrye, T., Mbada, C.E., et al.: Cost effectiveness of virtual reality game compared to clinic based Mckenzie extension therapy for chronic non-specific low back pain. Bull. Geodesique **16**(6), 601–609 (2022)
5. Lipchik, G.L., Nicholson, R.A., Penzien, D.B.: Allocation of patients to conditions in headache clinical trials: randomization, stratification, and treatment matching. Headache **45**(5)(498), 419–428 (2005)
6. Brown, B.W.J.: Designing for cancer clinical trials: selection of prognostic factors. Cancer Treat. Rep. **64**(2–3), 499–502 (1980)
7. Ross, S.M.: Introduction to Probability and Statistics for Engineers and Scientists, 6th edn. Academic Press, San Diego, CA (2020)
8. Morgado, L., Teodoro, F., Perdocoulis, T.: Manual de Metodos Estatísticos em Ciencias Biomedicas. Universidade de Trás os Montes e Alto Douro (UTAD, Vila Real, série Didáctica Cências Aplicadas **400** (2010). ISBN 978-972-669-018-2
9. Larsen, R., Marx, M.L.: Statistics, 5th edn. Prentice-Hall Inc. (2012)
10. Tamhane, A., Dunlop, D.: Statistics and Data Analysis. Prentice-Hall Inc., New Jersey (2000)
11. Thadewald, T., Buning, H.: Jarque-Bera test and its competitors for testing normality: a power comparison. Working Paper No.2004/9. Free University Berlin, School of Business & Economics, Berlin (2004). https://hdl.handle.net/10419/49919. Accessed 14 Apr 2025
12. Pocock, S.J.: Allocation of patients to treatment in clinical trials. Biometrics **35**(1), 183–197 (1979)
13. Pocock, S.J.: Clinical Trials, 5th edn. e-book. John Wiley & Sons, New Jersey (2013)
14. Atkinson, A.C., Biswas, A.: Bayesian adaptive biased-coin designs for bayesian adaptive biased-coin designs for clinical trials with normal responses. Biometrics **61**(1), 118–125 (2005)
15. Gail, M.H., Mark, S.D., Carroll, R.J., et al.: On design considerations and randomization based inference for community intervention trials. Stat. Med. **15**(11), 1069–1092 (1996)
16. Simon, R.: Restricted randomization designs in clinical trials. Biometrics **35**(2), 503–612 (1979)
17. Ling, S.X., Hobbs, B.P., Kaizer, A.M., Koopmeiners, J.S.: Calibrated dynamic borrowing using capping priors. J. Biopharm. Stat. **31**(6), 852–867 (2021)
18. Daniells, L., Mozgunov, P., Bedding, A., Jaki, T.: A comparison of bayesian information borrowing methods in basket trials and a novel proposal of modified exchangeability-nonexchangeability method. Stat. Med. **42**(24), 4392–4417 (2023)
19. Bruce, C.L., Iflaifel, M., Montgomery, A., et al.: Choosing and evaluating randomisation methods in clinical trials: a qualitative study. Clin. Trial **25**(1), 199 (2024)
20. Scott, D., Patterson, B.J.: Bioequivalence and Statistics in Clinical Pharmacology. Chapman and Hall/CRC, New York (2017)
21. Zeng, L., Qureshi, R., et al.: A registration of phase 3 crossover trials on clinicaltrials.gov. Trials **21**(613) (2020)

Computational Optimization and Applications (COA 2026)

Entropy Based Estimation of Risk Neutral Density Dynamics

Ana Fernandes[2], Rui Pascoal[1], and Ana M. Monteiro[1(✉)]

[1] Univ. Coimbra, Faculty of Economics, Av Dias da Silva 165, 3004-512 Coimbra, Portugal
ruiapsp@fe.uc.pt, ammm@fe.uc.pt

[2] Univ. Coimbra, CeBER, Faculty of Economics, Av Dias da Silva 165, 3004-512 Coimbra, Portugal
anasofiapfernandes@hotmail.com

Abstract. Derivative contracts constitute a fundamental component of financial markets, providing valuable insights into traders' forward-looking expectations. In particular, option contracts enable to retrieve the risk-neutral density (RND) of the underlying asset price. The maximum entropy approach offers a flexible framework for estimating the RND, especially in the presence of incomplete market information. Peter Buchen and Michael Kelly (1996), along with Leonidas Rompolis (2010), were among the first to apply the maximum entropy principle to RND estimation.

This study proposes a relative measure to capture the day-to-day evolution of changes in the RND by jointly employing maximum entropy and minimum cross-entropy methods. Moments based Rompolis's methodology is implemented. To address the ill-conditioning problem of the Buchen and Kelly approach we propose a strike selection procedure based on open interest data. The methodology is applied to both simulated data and market data from Microsoft, as well as the S&P 500 and VIX indices. Concerning market data, the proposed relative measure allows us to conclude that daily changes in the risk-neutral density estimates are not significant. (This study results from a master's dissertation archived in the Scientific Repository of the University of Coimbra [16].)

Keywords: Entropy · density estimation · financial derivative contracts · optimization

1 Introduction

Financial markets play a crucial role in the global economy, as market transactions incorporate traders' beliefs and expectations regarding future financial conditions. Among the wide range of financial instruments, derivatives stand out due to their diversity in form and purpose. Financial derivatives have been responsible, among other types of contracts, for a significant increase in trading volume.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 149–166, 2027.
https://doi.org/10.1007/978-3-032-30521-3_10

There are several types of derivatives, with futures, forwards, and options being among the most prominent. In particular, option contracts provide valuable information about investors' expectations concerning the price of the underlying asset at the contract's maturity date.

Cox and Ross [14] showed that the value of an option can be expressed as the discounted expected value of its future payoff under the risk-neutral probability distribution of the underlying asset price at maturity. Subsequently, Banz and Miller [7] as well as Breeden and Litzenberger [11], demonstrated that the second derivative of the option pricing function with respect to the strike price is proportional to the risk-neutral density (RND) in an arbitrage-free market.

Due to the richness of the information embedded in option prices, estimating the risk-neutral density at maturity serves several important purposes, including the pricing of illiquid and exotic options, as well as providing insights into investors' risk preferences.

A wide variety of methods has been proposed for estimating RNDs, which can be broadly classified into structural and non-structural approaches. Structural methods assume a stochastic process for the underlying asset and, in some cases, for its volatility (e.g., [9,10,19,24]). In contrast, non-structural methods directly approximate the RND function, potentially involving a large number of parameters. These methods can be further divided into parametric, non-parametric, and semi-nonparametric approaches. Parametric methods rely on a small number of parameters [5,23], whereas non-parametric methods allow for greater flexibility in the shape of the density by using a larger number of parameters [1–3,15,18,33]. Semi-nonparametric methods lie between these two extremes [12]. Some non-parametric approaches rely on estimating the option pricing function and its second derivative [11,14], while others are based on the principle of maximum entropy. Shannon [30] introduced the concept of entropy as a measure of missing information, distinct from the Boltzmann–Gibbs formulation in physics. The maximum entropy principle aims to derive probability distributions from limited information by imposing a set of constraints [20]. In this framework, the maximum entropy density (MED) is considered the "least biased" estimate, as it incorporates all available information without introducing unwarranted assumptions. When applied to financial markets, maximizing entropy leads to the most unbiased probability distribution consistent with observed option prices.

Buchen and Kelly [13] were among the first to apply the maximum entropy principle to recover the RND from European option prices. Stutzer [31] proposed an alternative approach that combines historical data on the underlying asset with entropy maximization to estimate the RND. Further applications include the estimation of volatility surfaces [4], as well as applications to interest rates and bond pricing [8,17].

One of the main objectives of this paper is to propose a measure for the day-to-day evolution of estimated RNDs. This measure is based on optimized entropy values, which arise as a byproduct of the estimation procedures. As discussed above, maximizing entropy subject to constraints derived from option prices allows the estimation of the RND for a given day. In contrast, cross-entropy

provides a measure of the proximity between a given distribution and a prior distribution (see Sect. 2), thus offering an absolute measure of the magnitude of distributional changes between consecutive days. Extensions involving generalized entropy measures, such as Tsallis and Rényi entropies, are explored in [22,32]. Entropy-based optimization methods thus provide a natural framework for constructing measures of the evolution of RNDs over time.

However, for meaningful comparisons over time, a relative measure is required. This is obtained by assessing how much farther the maximum entropy distribution at a given day is from the previous day's distribution, compared to the closest possible distribution consistent with current information. This closest distribution is obtained via minimum cross-entropy.

Rompolis [28], instead of using option prices directly as constraints, estimates the RND based on empirical risk-neutral moments of log-returns derived from option prices. Related formulations for extracting non-central moments are presented in [6,29]. Neri [25] and Schneider [26] estimate maximum entropy distributions using exotic options and apply them to interpolate implied volatilities.

The approach proposed in this paper addresses the issue of ill-conditioning by selecting relevant strike prices based on open interest, which serves as a proxy for the significance of each option contract. A novel empirical feature is identified: the magnitude of open interest influences the selection of relevant strikes and, consequently, the convergence of the estimation methods. This methodology allows focusing on the most informative option prices while mitigating numerical instability, particularly issues related to the ill-conditioning of the Hessian matrix. The quality of the proposed approach is evaluated using both simulated and empirical data. In the case of S&P 500 data, an interesting pattern is observed: open interest exhibits periodic peaks with a cycle of approximately 50.

The remainder of the paper is organized as follows. Section 2 presents the optimization problems associated with maximum entropy and minimum cross-entropy within the Buchen and Kelly framework, as well as the maximum entropy formulation in the Rompolis's approach. Accuracy metrics, including the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE), are also introduced. Section 3 applies these methods to simulated data generated from the Black–Scholes and Heston models, as well as to empirical data from the S&P 500, VIX, and Microsoft (MSFT). Measures of RND dynamics are computed and analyzed. Finally, Sect. 4 concludes the paper.

2 Maximum Entropy Density Estimation Methods Using Option Prices

The Principle of Maximum Entropy (PME) allows to select the distribution that maximizes the entropy measure with appropriate constraints.

Consider X a continuous random variable, with probability density function $f(.)$, representing either the price of a security at some future moment T or some price function such as its log-return prices, taking values in an interval I.

According to the maximum entropy principle, the estimation of the density $f(.)$ is obtained from the optimization problem,

$$maximize - \int_I h(f(x))dx \tag{1}$$

$$s.t. \quad \int_I f(x)dx = 1 \tag{2}$$

$$E[h_i(X)] = a_i \tag{3}$$

$$f(x) \geq 0, \quad \forall x \in I \tag{4}$$

where h is a proper choice, in this paper it is considered the Shannon entropy,

$$-\int_I f(x)log(f(x))dx,$$

$I = [0, K[,\ K > 0$, h_i are proper functions.

Buchen and Kelly [13] defined an optimization problem related to the RND estimation obtained from option prices, and based on Shannon entropy as:

$$maximize \quad -\int_0^{+\infty} f(x)log(f(x))dx \tag{5}$$

$$s.t. \quad \int_0^{+\infty} f(x)dx = 1 \tag{6}$$

$$C_i = E[e^{-r\tau}(X - K_i)^+] = \int_0^{+\infty} e^{-r\tau} f(x)(x - K_i)^+ dx,\ i = 1 \ldots, nc \tag{7}$$

$$P_j = E[e^{-r\tau}(K_j - X)^+] = \int_0^{+\infty} e^{-r\tau} f(x)(K_j - x)^+ dx,\ i = 1 \ldots, np \tag{8}$$

where X is the underlying asset price, r is the risk-free interest rate and C_i, $i = 1 \ldots, nc$, and P_j, $j = 1, \ldots, np$, are the derivatives prices, respectively the call and put option prices, with maturity T, $z^+ = max(0, z)$, K_i are the call strike prices, K_j are the put strike prices, nc (resp. np) is the number of calls (resp. puts).

Rompolis [28] optimization problem has the same objective function as in [13], the main difference between them lies in the constraints, [28] uses risk neutral non-central moments, expressed in terms of option prices as constraints. By considering $X = \log\left(\frac{S_T}{S_t}\right)$, where S_t is the underlying asset price, [28] defined an optimization problem related to the RND estimation obtained from option

prices as

$$maximize - \int_D f(x)log(f(x))dx \tag{9}$$

$$s.t. \quad \int_D f(x)dx = 1 \tag{10}$$

$$\int_D x^i f(x)dx = \mu_i,\ i = 1\ldots, m \tag{11}$$

$$f(x) \geq 0, \quad \forall x \in I \tag{12}$$

where $D = \left[\log\left(\frac{K_0}{S_t}\right), \log\left(\frac{K_\infty}{S_t}\right)\right]$, K_0 and K_∞ are the minimum and maximum strike prices given by data, respectively, μ_i denotes the ith order non-central moment of X and $\mu_0 = 1$. From constraint Eq. (11), for $i = 1, 2, 3, 4$, we may obtain the mean, the variance, the skewness, and the kurtosis coefficients, respectively.

Bakshi et al. [6] proposed a problem for recovering RND from option prices through non-central moments, where the risk-neutral moments are represented as follows:

$$\mu_1 = e^{r\tau}\left(1 - \int_{S_t}^{+\infty} \frac{1}{K^2} C_t(\tau, K)dK - \int_0^{S_t} \frac{1}{K^2} P_t(\tau, K)dK\right) - 1$$

$$\mu_i = e^{r\tau}\left(\int_{S_t}^{+\infty} \frac{i}{K^2}\left[log\left(\frac{K}{S_t}\right)^{i-2}\right]\left[i - 1 - log\left(\frac{K}{S_t}\right)\right] C_t(\tau, K)dK + \right. \tag{13}$$

$$\left. \int_0^{S_t} \frac{i}{K^2}\left[log\left(\frac{K}{S_t}\right)\right]^{i-2}\left[i - 1 - log\left(\frac{K}{S_t}\right)\right] P_t(\tau, K)dK\right),$$

where $C_t(\tau, K)$ and $P_t(\tau, K)$ denote the price of a European call and put at time t, respectively.

Rockinger and Jondeau [27] proposed an equivalent unconstrained minimization problem

$$minimize\ Z(\lambda_1, \ldots, \lambda_n) = \int_D exp\left(\sum_{i=1}^{m} \lambda_i(x^i - \mu_i)\right), \tag{14}$$

$\lambda = (\lambda_1, \ldots, \lambda_n)'$ represent the Lagrange multipliers. The solution of the above problem can be rewritten as:

$$f(x) = \frac{1}{Z(\lambda_1, \ldots, \lambda_n)} exp\left(\sum_{i=1}^{m} \lambda_i(x^i - \mu_i)\right). \tag{15}$$

Given up to 4 constraints related to four moments, there is not a closed form density, and therefore problem (9)–(12) is solved only through a numerical method.

The Principle of Minimum Cross-Entropy (PMCE) (Kullback-Leibler) is a further extension of the principle of maximum entropy [13,21]. The cross-entropy is a measure of the distance between two densities f and g, [13]. The PMCE aim is to find a density f which, out of all the densities satisfying the given constraints, is closest to g. The objective function to be minimized is defined by:

$$H_C(f,g) = \int_0^{+\infty} f(x) log\left(\frac{f(x)}{g(x)}\right) \tag{16}$$

where f and g are densities, and in this context g is a prior density. The PME maximizes uncertainty while PMCE minimizes a probabilistic distance to a given distribution. Maximizing entropy is equivalent to minimizing cross-entropy with respect to a uniform distribution.

Buchen and Kelly [13] presented the optimization problem that estimates the RND function from option prices, for a given prior distribution g:

$$\begin{aligned} &minimize - \int_0^{+\infty} f(x) log\left(\frac{f(x)}{g(x)}\right) dx \\ &s.t. \qquad \int_0^{+\infty} f(x)dx = 1 \\ &\qquad E[e^{-r\tau}(X-K_i)^+] = \int_0^{+\infty} f(x)e^{-r\tau}(x-K_i)^+ dx = C_i,\ i = 1\ldots, n_c \\ &\qquad E[e^{-r\tau}(K_j-X)^+] = \int_0^{+\infty} f(x)e^{-r\tau}(K_j-x)^+ dx = P_j,\ i = 1\ldots, n_p \\ &\qquad f(x) \geq 0, \quad \forall x \in [0,+\infty[. \end{aligned} \tag{17}$$

As stated before, we will evaluate the quality of these achievements. One of the purposes of this paper is to overcome the drawback in convergence, pointed out by [13], by choosing as relevant strikes those having high open interests.

Accuracy Metrics

To evaluate the outcome of the models we employ the Mean Absolute Error (MAE) that expresses the average absolute difference between estimates and real values, and Root Mean Square Error (RMSE) that measures the square root of the average squared difference between estimates and real values. Considering $\hat{V}_i$ as estimated prices and V_i as observed prices we have:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{V}_i - V_i)^2}{n}},$$

$$MAE = \frac{\sum_{i=1}^{n}|\hat{V}_i - V_i|}{n}.$$

3 Numerical Experiments

The entropy optimization problems will be tested by using simulated and market option prices. The former are simulated from Black-Scholes-Merton and Heston models. Market option prices on S&P500 and VIX indexes and Microsoft option data are collected from Chicago Board on Options Exchange (CBOE) market.

The optimization problems related to the Problem of Maximum Entropy (PME) are given by (5)–(8) ([13] approach) and (9)–(12) ([28] approach). The optimization problem related to the Principle of Minimum Cross-Entropy problem is given by (17).

Rompolis [28] method was implemented for $m \in \{2, 4\}$ and was designated by numerical maximum entropy density estimation based on risk-neutral moments (MED-RNM-N). For $m = 2$ we can obtain the analytical solution of the optimization problem, therefore, this approach is denoted by analytical maximum entropy density estimation based on risk-neutral moments (MED-RNM-A). Since the approach for $m = 4$ gives poor results, we choose to compare the analytical solution with the numerical solution for $m = 2$.

A simulation analysis for BSM and Heston data is performed using the routines *blsprice* and *optByHestonNI*, respectively, from Financial Toolbox of Matlab. For implementing Buchen and Kelly's method we used the CVX routine which is a matlab-based program for convex optimization, complemented with new built functions.

3.1 Simulation Data Black-Scholes Model

The stock price model proposed by [10] and [24] assume that the underlying asset follows a geometric Brownian motion,

$$dS_t = S_t \mu dt + S_t \sigma dW_t,$$

where μ and σ are constants and denote the expected return and the volatility, respectively, S_t is the underlying asset price at time t and W_t a standard Brownian motion. Applying Itô Lemma allows to specify that the probability distribution of $log(S_T)$, $T > t$, is given by

$$log(S_T) \sim N\left(log(S_t) + (\mu - \frac{1}{2}\sigma^2)\tau, \sigma^2\tau\right),$$

where $\tau = T - t$ and $N(\alpha, \beta^2)$ denotes a normal distribution with mean α and variance β^2. Therefore S_T follows a log-normal distribution.

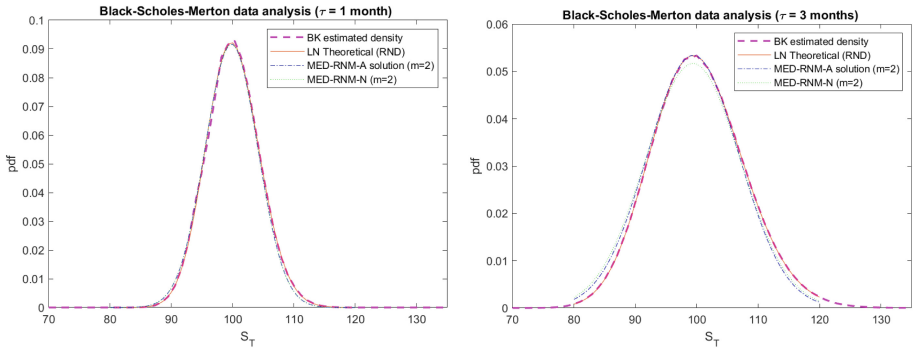
In order to generate BSM option prices, the chosen model parameters are: the current underlying asset price, $S_t = 100$, the risk-free interest rate, $r = 0.01$, which is the return under risk neutral distribution and the volatility, $\sigma = 0.15$. We considered 21 strike prices, K, equally spaced between 80 and 120, and two times to maturity $\tau \in \{\frac{1}{12}, \frac{3}{12}\}$. Therefore, we generated 42 option prices (21 calls and 21 puts), for each maturity.

Due to different liquidity of the option contracts and transaction costs, the simulated call prices were perturbed. To reproduce market behavior, random noise is added to the simulated prices:

$$C_i^+ = (1 + 0.005\xi_i)C_i \tag{18}$$

$$P_i^+ = (1 + 0.005\eta_i)P_i \tag{19}$$

$\xi \sim N(0,1)$, $\eta \sim N(0,1)$, and C_i, P_i are the theoretical call and put prices computed by BSM formula. The multiplicative form of (18)–(19) allows to mimic market heterocedasticity.



(a) (b)

Fig. 1. Theoretical (lognormal) RND and estimated RNDs from BSM data - Buchen and Kelly, and Rompolis methods - 1 and 3 months

Table 1. RMSE and MAE for estimated call and put option prices (BSM data)

	Buchen and Kelly method	MDN-RNM-A method	MDN-RNM-N method (m = 2)
RMSE (one month)	3.2695×10^{-4}	1.2×10^{-2}	1.3×10^{-2}
RMSE (three months)	1.5307×10^{-4}	8.8717×10^{-2}	8.8749×10^{-2}
MAE (one month)	1.0596×10^{-4}	8.817×10^{-3}	8.817×10^{-3}
MAE (three months)	4.1365×10^{-5}	1.3×10^{-2}	1.6×10^{-2}

By analyzing the RMSE and MAE values in Table 1, we observe that the estimated values are accurate, which supports the validity of the proposed estimation methods for theoretical data. The results appear to be more accurate for Buchen and Kelly's method than for Rompolis's method. In Fig. 1, the theoretical and estimated densities are plotted together. Overall, the estimation

shows a good fit, with better performance for Buchen and Kelly's method. However, some issues arise with the Rompolis's method. We observe that empirical RND moments may not capture all the relevant information contained in option prices. In particular, different behaviors are observed across distinct regions of the density support. In this regard, Buchen and Kelly's method is more flexible, as it captures different behaviors between each pair of consecutive strikes.

3.2 Heston Data Analysis

The effectiveness of the Buchen and Kelly's method is also analyzed through estimation of RNDs for Heston model, [19]. He assumed that volatility varies stochastically. The dynamic model has two stochastic components, the underlying asset price S_t and its variance V_t,

$$dS_t = S_t \mu dt + S_t \sqrt{V_t} dWt, 1, \tag{20}$$

$$dV_t = \mathcal{K}(\theta - V_t)dt + \sigma \sqrt{V_t} dWt, 2, \tag{21}$$

where $dW_{t,1}$ and $dW_{t,2}$ are standard Brownian motions with correlation coefficient ρ, $\mathcal{K}$ represents the volatility mean reversion speed for V_t, θ denotes the long-term mean variance, and σ is the volatility of the volatility process.

The chosen parameters values are: risk-free interest rate, $r = 0.05$; initial variance for the underlying asset, $V_t = 0.026$; volatility mean reversion speed for V_t, $\mathcal{K} = 3.93$; long-term mean of variance, $\theta = 0.01$; volatility of volatility $\sigma = 0.04$; the maturity, $\tau = \frac{1}{12}$ and the underlying asset price, $S_t = 1080$. Also, it is considered 12 strike prices equally spaced between 820 and 1080, corresponding to 12 calls and 12 put prices.

Different correlation coefficients are considered, $\rho \in \{-0.52, 0, 0.52\}$. To reproduce the market's behavior, random noise is added to theoretical option prices. Hence, we replaced these prices as follows, $C_i^* = (1 + 0.005 \cdot \xi_i) \cdot C_i$ and $P_j^* = (1 + 0.005 \cdot \eta_j) \cdot P_j$, with $\xi \sim \mathcal{N}(0, 1)$ and $\eta \sim \mathcal{N}(0, 1)$.

Table 2 compares theoretical and estimated densities, it is observed that for all ρ values, the estimation is overall well fitted. Nonetheless, the obtained estimates present worst adjustment than for BSM data (though not so clear for MAE measure), even without noise. Figure 2 presents that for $\rho = 0.52$ the skewness is positive, which means that the right tail is heavier, it shows also that by perturbing option prices, we get worse estimates.

Table 2. RMSE and MAE for estimated call and put option prices (Heston data without noise)

ρ	RMSE	MAE
-0.52	1.1×10^{-2}	7.6860×10^{-4}
0	1.1×10^{-2}	7.6434×10^{-4}
0.52	1.3×10^{-2}	7.7524×10^{-4}

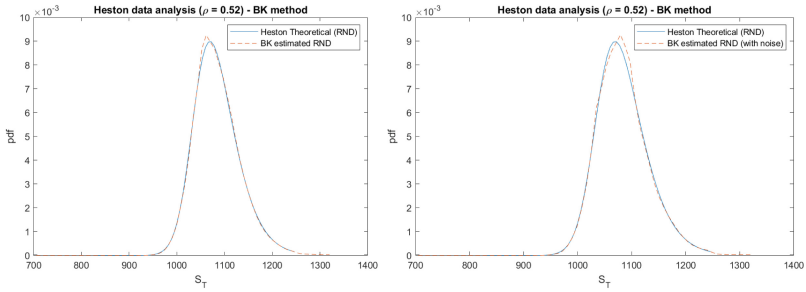


Fig. 2. Heston theoretical RND and estimated RNDs from Heston data (with and without noise)

3.3 Market Data Analysis

Risk neutral densities are recovered using option prices from $S\&P500$ and VIX indexes and from Microsoft data. These data were obtained from CBOE exchange market. We choose Buchen and Kelly's method to evaluate the densities.

Dataset Description. For each sample, we considered the arithmetic mean of bid and ask option prices. The interest rate data is collected from the Central Bank of the Federal Reserve System of the United States of America. To ensure the quality of the data, we selected call and put option prices with larger open interests since they can possibly better represent the market's trader expectations. For $S\&P500$ data, we consider a sequence of strikes with a constant difference of 50 between consecutive ones, since these were the ones with remarkably higher open interest values, for the period considered. When considering other strike prices, we didn't obtain better results and in some cases not even convergence. We stress that this is a feature that strongly improves the quality and feasibility of Buchen and Kelly's estimation.

We consider three datasets from the $S\&P500$ index. The first dataset collects data from November 1st to 12th, 2021, with maturity on December 17th, 2021. The second dataset collects data from January 10th to 14th, 2022, with maturity on February 18th, 2022, this dataset also contains data from VIX from January 10th to 14th, 2022, with maturity on February 16th, 2022. The third dataset collects data from November 1st, 2021, with maturity on January 21st, 2022, $S_t = 4614$, and $r = 0.05\%$. We estimated maximum entropy densities for all datasets and minimum cross-entropy density for the first dataset. Table 3 present some information about first and second datasets regarding S&P500 and VIX, respectively.

Microsoft dataset includes data from December 1st, 2021, with maturity on June 17th, 2022. The underlying asset price, S_t, is 330 and the interest rate, r, is 0.001. Since Microsoft option data are American style options, we selected only out-of-the money options.

Table 3. Datasets

First dataset - $S\&P500$				Second dataset - VIX			
Day	τ (days)	S_t	$r(\%)$	Day	τ (days)	S_t	$r(\%)$
November 1st	46	4613	0.08	January 10th	37	19.4	0.05
November 2nd	45	4630	0.06	January 11th	36	18.41	0.04
November 3rd	44	4660	0.07	January 12th	35	17.62	0.04
November 4th	43	4680	0.05	January 13th	34	20.31	0.05
November 5th	42	4697	0.06	January 14th	33	19.19	0.05
November 8th	39	4701	0.05				
November 9th	38	4685	0.05				
November 10th	37	4646	0.06				
November 11th	36	4649	0.06				
November 12th	35	4682	0.05				

Estimated RND Results. Considering the first dataset for S&P500 data, Figs. 3, 4 present the estimated daily RND.

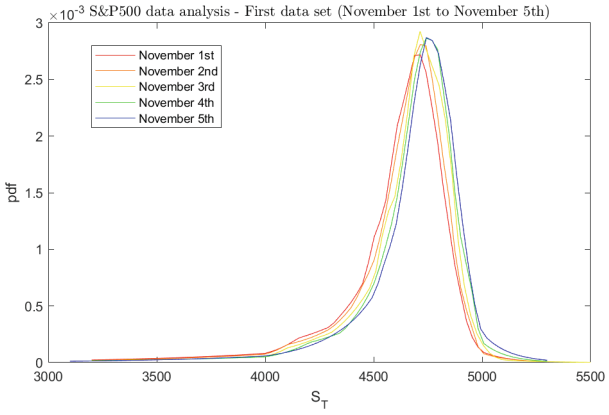


Fig. 3. Estimated RNDs from $S\&P500$ (first dataset) - November 1st to November 5th

As for Table 4, we present the statistical parameters for estimated RNDs. The mean for the density estimates is situated between 4604 and 4693. Remark that, in general, the mean values are near the current value (Table 3).

Considering Figs. 3, 4, and Table 4 we observe that skewness is negative, which means that left tail is heavier, and therefore the probability of strongly negative returns is higher than for positive returns. The kurtosis coefficient is greater than 3 (leptokurtic distribution), that suggest there may be a large probability of extreme values occurrence on both sides.

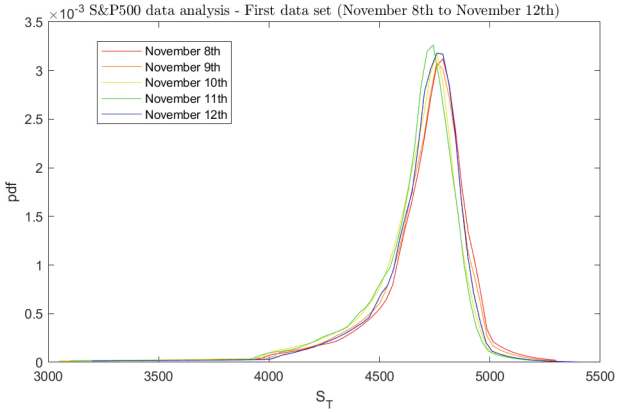


Fig. 4. Estimated RNDs from $S\&P500$ (first dataset) - November 8th to November 12th

Table 4. Statistical parameters for RND estimate (first dataset) - $S\&P500$

Day	Mean	Variance	Skewness	Kurtosis
November 1st	4604	6832	−2.0395	9.0948
November 2nd	4621	6467	−2.1058	9.5512
November 3rd	4651	5835	−2.1755	10.5705
November 4th	4672	5794	−2.2311	11.0139
November 5th	4685	6332	−2.2628	11.2655
November 8th	4693	6230	−2.4338	12.7395
November 9th	4676	6308	−2.3495	11.8706
November 10th	4642	6567	−2.3242	11.3210
November 11th	4645	5754	−2.3119	11.7340
November 12th	4682	4973	−2.3259	12.2705

There seems to be a relationship between skewness and kurtosis. When skewness goes up (resp. down) in absolute value, so does the kurtosis.

Considering the estimation applied to $S\&P500$ data for the second dataset, Table 5, we observe that as in the previous case, the skewness is negative, and the skewness and kurtosis increase when the mean increases. Negative skewness may indicate that investors are somewhat pessimistic.

As for the estimation applied to VIX data, Table 6 and Fig. 5, we observe that the asymmetry is positive. There seems to be a nearly inverse relationship between the asymmetry obtained for the $S\&P500$ and for the VIX, see Tables 5 and 6. When skewness increases in absolute value for $S\&P500$ data, it decreases for VIX data, and vice versa, in general, as it is expected. In addition, we can state that for $S\&P500$ data, the main interest is the estimation of the left tail

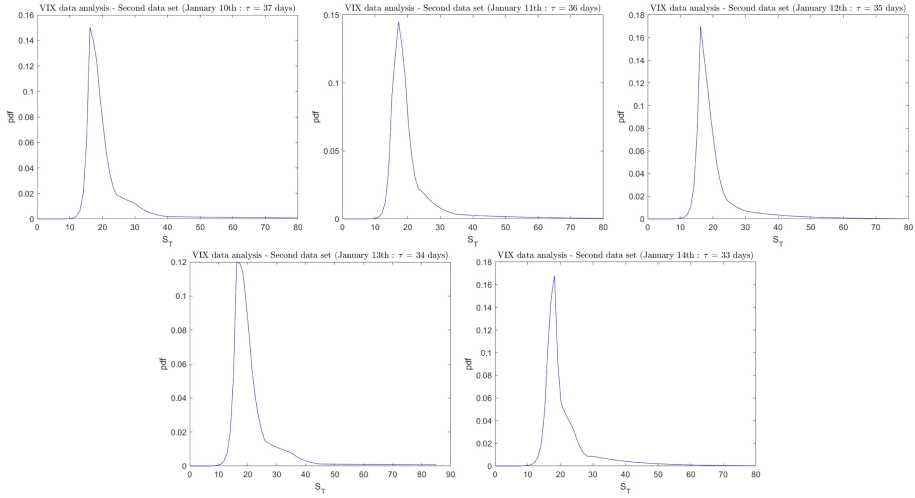


Fig. 5. Estimated RNDs from VIX (second dataset)

Table 5. Statistical parameters for estimated RNDs (second dataset) - $S\&P500$

Day	Mean	Variance	Skewness	Kurtosis
January 10th	4666	7417	−1.9786	8.9133
January 11th	4709	6765	−2.0481	9.4916
January 12th	4725	6179	−2.0719	9.8647
January 13th	4656	7464	−1.8074	7.6025
January 14th	4663	6584	−1.8186	7.9930

associated with more risk, whereas, for VIX data, it is the right one associated with more volatility, see Figure 5.

Considering Fig. 6 and Table 7, S&P500 - third dataset, we observe that just as in the first dataset, the skewness is negative, and the kurtosis is high. As for Fig. 7 and Table 7, MSFT, we observe that just as for the previous estimated RND, the skewness is negative, and the kurtosis is high, but both seem to be more limited than for S&P500.

Comparison Between Minimum Cross-Entropy and Maximum Entropy Densities. In this paper it is proposed an indicator that allows us to compare each RND estimate to that of the previous day, analyzing if there was a substantial variation. This indicator, $M(H_C)$, is defined as follows:

$$M(H_C)_t = \frac{H_C(f_t, f_{t-1})}{H_{min}} = 1 + \frac{H_C(f_t, f_{t-1}) - H_{min}}{H_{min}} \tag{22}$$

Table 6. Statistical parameters for estimated RNDs - VIX

Day	Mean	Variance	Skewness	Kurtosis
January 10th	21.68	91	3.2825	15.4297
January 11th	21.10	84	3.2474	15.3343
January 12th	20.85	72	3.0919	14.4984
January 13th	22.15	94	3.3166	16.1484
January 14th	21.85	80	2.8095	12.4113

Table 7. Statistical parameters for estimated RND (third dataset) - $S\&P500$ and $MSFT$

	Mean	Variance	Skewness	Kurtosis
Third dataset- $S\&P500$	4661	1265	−1.7981	7.2333
$MSFT$	330	5553	−0.3525	4.2285

where f_t (resp. f_{t-1}) represents the maximum entropy density, estimated at day t (resp. $t-1$); $H(f_t, f_{t-1})$ denotes the cross-entropy between f_t and f_{t-1}; H_{min} denotes the cross-entropy between f_t^* and f_{t-1} defining f_t^* as the density closest to f_{t-1}, given time t, that is, the density obtained by fitting the minimum cross-entropy with respect to f_{t-1}, subject to the constraints at time t.

If $M(H_C)$ is very close to 1, there is little change with respect to the estimated density of the previous day since the estimated density for day t does not have a cross-entropy with day $t-1$ density much higher than the possible minimum one. Therefore $M(H_C)$ measure can serve as an effective diagnostic tool for identifying days with the most significant changes in the market sentiment, allowing for the subsequent identification of explanations based, for example, on unexpected events. It is also worth examining which density zones exhibit the greatest divergence between H_C and H_{min} on such days.

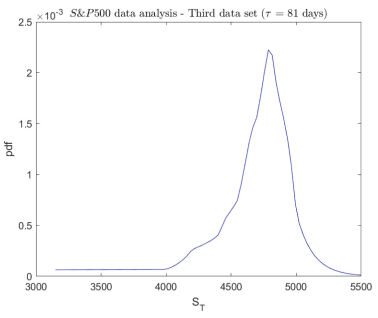


Fig. 6. Estimated RND from $S\&P500$ (third dataset)

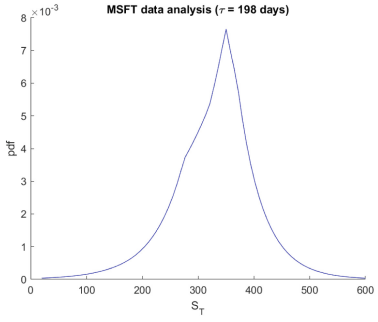


Fig. 7. Estimated RNDs from $MSFT$

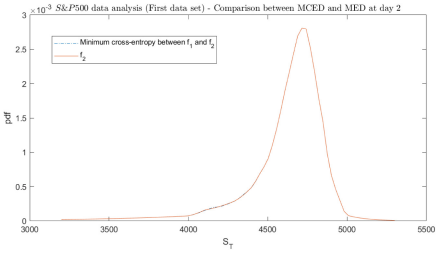


Fig. 8. $S\&P500$ data analysis (First dataset) - Comparison between MCED and MED at day November 2nd.

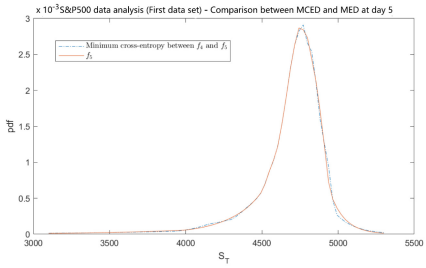


Fig. 9. $S\&P500$ (First dataset) - Comparison between MCED and MED at day November 5th

Considering Fig. 8, it is illustrated the comparison between the minimum cross-entropy density (MCED) and the maximum entropy density (MED) estimates, for the first dataset, November 2nd. The estimates obtained for MCED are similar to those obtained for MED, this may signify that there is little difference between the current density and the density estimate of the previous day.

Through Table 8 it is possible to perceive that the highest ratio, $M(H_C)$, was on November 11th, followed by November 5th. Figure 9, corresponding to November 5th, shows that MCED estimate does not adjust so well to MED estimate as for November 2nd (Fig. 8). It seems that the ratio reflects well changes in the estimated densities, over time. It would be interesting to test the behavior of the ratio in case of more drastic evolution in densities. The highest H_{min} value was for November 4th and the lowest value for November 9th.

Table 8. Indicator $M(H_C)$

	H_C	H_{min}	$M(H_C)$
November 2nd (prior: November 1st)	0.0055	0.0054	1.0031
November 3rd (prior: November 2nd)	0.2851	0.2847	1.0012
November 4th (prior: November 3rd)	0.7332	0.7331	1.0012
November 5th (prior: November 4th)	0.0359	0.0354	1.0673
November 8th (prior: November 5th)	0.0119	0.0119	1.0057
November 9th (prior: November 8th)	0.0051	0.0050	1.0260
November 10th (prior: November 9th)	0.0211	0.0208	1.0179
November 11th (prior: November 10th)	0.0146	0.0134	1.0934
November 12th (prior: November 11th)	0.1493	0.1486	1.0044

4 Conclusions

Option prices can provide relevant information to evaluate the traders' expectations and risk preferences. From option data, it is then possible to estimate the RND, which could reflect the market sentiment that traders attribute to the future price of the underlying asset. This work applies entropy optimization methods based on PME, Buchen and Kelly and Rompolis methods, and on PMCE for Buchen and Kelly method. For this purpose, we used simulated data from BSM and Heston models, and market data on S&P500, VIX, and MSFT.

Furthermore, innovations were developed for implementing these methods for market data. First, a selection of strikes was performed in order to operationalize the Buchen and Kelly method. Second, a new measure was conceived to analyse a short period of time, and perceive if there was a substantial variation in RND at each day regarding the previous day.

Through graphical analysis for the simulated data, we highlight that both methods produce valid risk-neutral density estimates. We obtained proper RND estimates, always with non-negative values and with expected tail behavior. Regarding the market data, the estimated RNDs verify the expected behavior. It is possible to observe for S&P500 and MSFT datasets that investors seem to be pessimistic about the market since skewness is negative. In addition, the kurtosis is high, which may signify that the risk is high. Regarding VIX data, the skewness is positive in opposite to the results from S&P500 and Microsoft. The right tail is the main interest of the estimation since it can reflects traders' concerns with high price flutuations, namely the negative ones.

Acknowledgments. This work has been funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., through project UID/05037/2025 with DOI https://doi.org/10.54499/UID/05037/2025, and through the project "Complex Risk Management in the Big Data Regime" – PTDC/MAT–APL/1286/2021.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aït-Sahalia, Y., Duarte, J.: Nonparametric option pricing under shape restrictions. J. Econometrics **116**(1), 9–47 (2003)
2. Aït-Sahalia, Y., Lo, A.W.: Nonparametric estimation of state-price densities implicit in financial asset prices. J. Financ. **53**(2), 499–547 (1998)
3. Ardakani, O.M.: Option pricing with maximum entropy densities: the inclusion of higher-order moments. J. Futur. Mark. **42**(10), 1821–1836 (2022)
4. Avellaneda, M.: Minimum-relative-entropy calibration of asset-pricing models. Int. J. Theor. Appl. Financ. **1**(04), 447–472 (1998)
5. Bahra, B.: Implied risk-neutral probability density functions from option prices: theory and application. Bank of England Working Paper (1997)
6. Bakshi, G., Kapadia, N., Madan, D.: Stock return characteristics, skew laws, and the differential pricing of individual equity options. Rev. Financ. Stud. **16**(1), 101–143 (2003)

7. Banz, R.W., Miller, M.H.: Prices for state-contingent claims: some estimates and applications. J. Bus. 653–672 (1978)
8. Bariviera, A.F., Martín, M.T., Plastino, A., Vampa, V.: Libor troubles: anomalous movements detection based on maximum entropy. Phys. A **449**, 401–407 (2016)
9. Bates, D.S.: Jumps and stochastic volatility: exchange rate processes implicit in deutsche mark options. Rev. Financ. Stud. **9**(1), 69–107 (1996)
10. Black, F., Scholes, M.: The pricing of options and corporate liabilities. J. Polit. Econ. **81**(3), 637–654 (1973)
11. Breeden, D.T., Litzenberger, R.H.: Prices of state-contingent claims implicit in option prices. J. Bus. 621–651 (1978)
12. Bu, R., Hadri, K.: Estimating option implied risk-neutral densities using spline and hypergeometric functions. Economet. J. **10**(2), 216–244 (2007)
13. Buchen, P.W., Kelly, M.: The maximum entropy distribution of an asset inferred from option prices. J. Financ. Quant. Anal. **31**(1), 143–159 (1996)
14. Cox, J.C., Ross, S.A.: The valuation of options for alternative stochastic processes. J. Financ. Econ. **3**(1–2), 145–166 (1976)
15. Dalderop, J.: Nonparametric filtering of conditional state-price densities. J. Econometrics **214**(1), 295–325 (2020)
16. Fernandes, A.: Analysis of investors expectations using entropy optimization models. Master's thesis, University of Coimbra (2022). https://estudogeral.uc.pt/handle/10316/100098
17. Gzyl, H., Mayoral, S.: Determination of zero-coupon and spot rates from treasury data by maximum entropy methods. Phys. A **456**, 38–50 (2016)
18. Härdle, W., Hlávka, Z.: Dynamics of state price densities. J. Econometrics **150**(1), 1–15 (2009)
19. Heston, S.L.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**(2), 327–343 (1993)
20. Jaynes, E.T.: Concentration of distributions at entropy maxima. ET Jaynes: Papers on Probability, Statistics and Statistical Physics p. 315 (1979)
21. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
22. Malhotra, G., Srivastava, R., Taneja, H.: Calibration of the risk-neutral density function by maximization of a two-parameter entropy. Phys. A **513**, 45–54 (2019)
23. Melick, W.R., Thomas, C.P.: Recovering an asset's implied pdf from option prices: an application to crude oil during the gulf crisis. J. Financ. Quant. Anal. **32**(1), 91–115 (1997)
24. Merton, R.C.: Theory of rational option pricing. Bell J. Econ. Manage. Sci. 141–183 (1973)
25. Neri, C., Schneider, L.: Maximum entropy distributions inferred from option portfolios on an asset. Finance Stochast. **16**, 293–318 (2012)
26. Neri, C., Schneider, L.: A family of maximum entropy densities matching call option prices. Appl. Math. Financ. **20**(6), 548–577 (2013)
27. Rockinger, M., Jondeau, E.: Entropy densities with an application to autoregressive conditional skewness and kurtosis. J. Econometrics **106**(1), 119–142 (2002)
28. Rompolis, L.S.: Retrieving risk neutral densities from European option prices based on the principle of maximum entropy. J. Empir. Financ. **17**(5), 918–937 (2010)
29. Rompolis, L.S., Tzavalis, E.: Recovering risk neutral densities from option prices: a new approach. J. Financ. Quant. Anal. **43**(4), 1037–1053 (2008)
30. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Technical J. **27**(3), 379–423 (1948)

31. Stutzer, M.: A simple nonparametric approach to derivative security valuation. J. Financ. **51**(5), 1633–1652 (1996)
32. Tanak, A.K., Borzadaran, G.M., Ahmadi, J.: Maximum Tsallis entropy with generalized Gini and Gini mean difference indices constraints. Phys. A **471**, 554–560 (2017)
33. Yatchew, A., Härdle, W.: Nonparametric state price density estimation using constrained least squares and the bootstrap. J. Econometrics **133**(2), 579–599 (2006)

Predicting Consumer Acceptance of Gamification Strategies: A Machine Learning Approach

Patrícia Santos[1], Cristela Bairrada[3], and Joana Dias[2(✉)]

[1] Faculty of Economics, University of Coimbra, Coimbra, Portugal
[2] Faculty of Economics and INESC Coimbra, University of Coimbra, Coimbra, Portugal
joana@fe.uc.pt
[3] Faculty of Economics and CeBER, University of Coimbra, Coimbra, Portugal
cristela.bairrada@uc.pt

Abstract. Gamification is currently a powerful marketing strategy that contributes to higher consumer engagement and brand loyalty. However, this strategy comes with significant implementation costs and potential high failure rates, so a careful market assessment should be carried out before investment. This work addresses a critical gap that currently exists by developing a machine learning-based predictive model to identify consumers that are most likely to adopt gamification. Grounded on the theory of consumption values and innovation adoption literature, three classification algorithms (Random Forests, Support Vector Machines, and MARS) were tested on data collected from consumers. Results demonstrate that Random Forests achieve 84% prediction accuracy (AUC: 0.91) using just three key variables: functional value, emotional value, and consumer innovativeness. These results provide managers with a simple and practical tool for market segmentation and targeted gamification deployment, contributing to the maximization of return on investment while minimizing risk, since it will prevent the adoption of a strategy that will not have the desired impact among consumers. The decision tree model that was built allows for a clear interpretation of decision rules. Practitioners will be able to quickly assess whether gamification aligns with specific customer segments and why.

Keywords: gamification · machine learning · consumer adoption · market segmentation · classification algorithms

1 Introduction

Gamification can be defined as the application of game design elements in non-game contexts. It has transformed significantly digital marketing strategies over the past decade and it continues to influence and impact current practices [1]. Some reasons that justify the growing popularity of gamification can be identified: the impressive increase of digital platforms, the proliferation of mobile technologies, and significant changes in consumer behavior, especially in the post-pandemic period [2]. The global gamification market, valued at approximately $36 billion in 2026, is projected to reach $112

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 167–181, 2027.
https://doi.org/10.1007/978-3-032-30521-3_11

billion by 2031, reflecting its huge growth across e-commerce and digital platforms [3]. The mechanisms that make gamification appealing are already well-documented: by incorporating points, badges, leaderboards, challenges, and narrative elements, for instance, organizations successfully increase consumer engagement (emotional, cognitive, and behavioral), improve brand loyalty, increase purchase intentions, and increase the likelihood of positive word-of-mouth [4–6].

Although it has been demonstrated that gamification clearly brings benefits, it is also important to acknowledge that its implementation can bring significant challenges. Firstly, it requires substantial investment in financial, human, and technological resources. Effective implementation of gamification involves high associated costs, similar to other marketing actions [7], such as developing websites or mobile applications [8], as well as development time, effort, expertise, continuous maintenance and some adjustments to match the brand's identity [7, 9]. Secondly, empirical evidence has shown that gamification effectiveness is far being guaranteed: 36% of consumers report that gamification does not influence purchasing decisions, and only 17% indicate that gamification captures their attention and increases purchase likelihood [10]. Thirdly, consumer responses to gamification vary significantly across demographic and psychological segments, with individual characteristics such as age, digital profile, personal motivations, and consumption habits significantly influencing engagement levels [11, 12].

These findings identify a strategic dilemma for organizations with limited budgets: how can managers evaluate feasibility and potential adoption of gamification strategies before assigning resources to these strategies? Recent literature emphasizes that market segmentation and targeted positioning of gamified strategies are essential to ensure effectiveness [13]. However, to the best of the authors' knowledge, prior work has not provided practical tools that support rapid, accurate market assessment during the pre-implementation phase in this context.

This study fills this gap by developing and validating a machine learning-based classification model that predicts individual consumer willingness to adopt gamification. Our research addresses three managerial questions: (1) What consumer characteristics determine gamification acceptance? (2) How can companies segment clients to target gamification to high-probability adopters? (3) Can a minimal set of questions enable rapid assessment of new clients' gamification receptivity?

Using data from consumers and comparing three classification algorithms, we demonstrate that Random Forests achieve 84% prediction accuracy using only three predictor variables grounded in consumption value theory and innovation adoption literature: functional value, emotional value, and consumer innovativeness.

2 Theoretical Framework

Gamification's effectiveness stems from its capacity to activate intrinsic motivations - autonomy, competence, and relatedness - through game mechanics [14]. Incorporating these elements into otherwise only practical experiences allows organizations to create emotional appeal and promote continued consumer engagement. Marketing research has shown that gamification has the potential to improve brand attitude, brand equity, and long-term customer lifetime value [15]. However, these benefits are strongly dependent on a correct alignment between game mechanics and target audience preferences.

The theory of consumption values provides a robust framework for understanding different consumer responses to marketing innovations including gamification [16]. Consumption value theory states that individuals choose alternatives based on multiple value dimensions [17, 18]:

- **Functional Value** captures the perceived utility that comes from practical and utilitarian attributes, including quality, efficiency, reliability, ease of use, price, and accessibility. In the context of gamification, it reflects whether consumers understand these attributes as being advantageous to them, such as easy point accumulation, transparent reward structures, and efficient engagement mechanisms.
- **Emotional Value** represents the perception of feelings or affective states evoked by an alternative - pleasure, enjoyment, fun, entertainment, and playfulness [14]. Gamification's core promise is emotional value creation; thus, consumers with high emotional value orientation toward gaming elements should demonstrate greater adoption propensity.
- **Social Value** is related with the image that individuals want to project among their social groups through their choice behavior [19]. It represents a consumer's belief that the use of gamified systems will increase his/her social status, acceptance, or approval through mechanisms such as leaderboards, social comparison, or public achievements [20].
- **Epistemic Value** refers to the perceived utility derived from novelty, curiosity, and opportunities for learning or exploration offered by gamification [21]. Epistemic value is generated when game elements introduce new experiences, challenges, narratives, or interactive features that stimulate curiosity and provide users with opportunities to acquire new knowledge or skills.
- **Conditional Value** represents the situational utility of an alternative given specific contexts or circumstances. In gamification contexts, the value of game elements may vary depending on temporal, task-related, or environmental conditions.

In this study, multiple value dimensions are examined in order to identify the primary drivers that enable the prediction of whether a customer will engage in a gamification experience.

3 Research Methodology

3.1 Data Collection and Sample

The empirical analysis focuses on current and potential users of a company that provides services specifically designed for students, whose primary market segment is university students. The questionnaire was therefore administered to individuals enrolled in higher education, reflecting the firm's actual and prospective customer base. This sampling frame ensures that the resulting predictive model is aligned with the decisions faced by managers operating in a student-services environment, where digital literacy is high and gamification is a plausible strategic option [22].

A total of 1,450 printed questionnaires were distributed to students. The full questionnaire is presented in Appendix 1. The participants were instructed to fill in the questionnaires and deposit them in ballot boxes so they could be totally reassured the

answers would be anonymous. Of the 1,450 questionnaires distributed, 605 were collected. Questionnaires containing missing data were discarded being the final sample composed of 347 questionnaires that were used in the analysis. As for the respondents, 54.5% were women and 45.5% men, the average age being 23 years, and 65.4% of whom were full time students. Initially, the questionnaire invited respondents to read a gamification scenario and then answer to all questions made considering that gaming experience described. The scenario was created based on characteristic elements of gamification systems widely used on digital platforms, to ensure the plausibility and representativeness of the described experience. The variables hypothesized to influence willingness to engage in gamification were measured using previously developed and validated measurement scales, adapted to this context, namely: Trust [23], Positive and Negative Affect Schedule (PANAS) [24], consumption values [25, 26], technology readiness [27], impulsive personality trait [28], big five [29], experience with games [30] and intention to enter [31]. A five-point Likert scale was used, ranging from 1 "strongly disagree" to 5 "strongly agree".

The dataset was reduced to a subset of relevant attributes, focusing on key demographic and behavioral variables, including gender, age, income, and purchase frequency. These are categorical variables. To reduce dimensionality while preserving the key dimensions of the study, several composite variables were created by averaging the multiple questionnaire items used to evaluate each dimension. This approach generated summary measures for trust, functionality, emotional engagement, social interaction, epistemic curiosity, conditions, optimism, innovation, technology, insecurity, discomfort, gaming, neuroticism, extraversion, openness, empathy, conscientiousness, and impulsivity, capturing the underlying information of multiple correlated items in a single score. All numeric variables were standardized using z-score normalization. This ensured that each variable had a mean of zero and a standard deviation of one. The target variable was transformed into a binary target variable (classification problem) by coding as 1 if the corresponding response score was greater than 3, indicating acceptance and 0 otherwise. The sample exhibited initial class imbalance: 191 records with target = 1 (engaging) and 156 with target = 0 (non-engaging).

3.2 Experimental Design and Cross-Validation Protocol

To reduce bias caused by class imbalance and ensure robust estimation across multiple random samples, the following procedure was applied: (1) **Balanced Dataset Generation**: We generated datasets with all the 156 non-engaging (negative) cases and then randomly sampled 156 cases with positive outcomes from the pool of 191, creating balanced datasets with 312 observations each (156 per class). (2) **Iterative Replication**: This balancing procedure was repeated 100 times using different random samples to ensure that the results were not influenced by any single sampling instance. (3) **Leave-One-Out Cross-Validation**: For each balanced dataset iteration, we employed leave-one-out cross-validation (LOOCV): each observation was excluded one at a time, models trained on the remaining 311 observations, and predictions generated for the held-out case. This procedure prevents overfitting and contributes to a more accurate assessment of generalization performance.

3.3 Classification Algorithms and Performance Metrics

We compared three state-of-the-art machine learning algorithms. All computational experiments were conducted using the R programming environment. Libraries used were: *rpart, earth, randomForest, pROC, caret, e1071.*

- **Random Forests** (RF) construct multiple decision trees through bootstrap sampling and random feature selection, aggregating predictions via majority vote. This ensemble approach reduces variance and mitigates overfitting while handling non-linear relationships and feature interactions effectively [32].
- **Support Vector Machines** (SVM) find optimal hyperplanes to separate classes in high-dimensional feature space, with kernel functions enabling non-linear decision boundaries. SVMs excel with smaller datasets and feature interaction patterns [33].
- **Multivariate Adaptive Regression Splines** (MARS) iteratively build piecewise linear models, capturing non-linear relationships through basis functions and interactions. MARS provides interpretable models while accommodating non-linearities [34].

Models were assessed considering **Accuracy** (percentage of correct predictions across balanced test sets) and Area Under the Receiver Operating Characteristic Curve (**AUC**) (measuring discrimination ability across threshold values; perfect predictor = 100%, random predictor = 50% on balanced data). Variable importance rankings were computed for each algorithm iteration, with average importance scores across 100 iterations providing robust estimates.

4 Results

4.1 Full-Feature Model Comparison

RF demonstrated the strongest predictive performance across both metrics (Table 1). Mean accuracy reached 83% (SD = 0.8%, range: 81–85%), while mean AUC was 91% (SD = 0.5%, range: 90–93%). Variable importance analysis identified six key predictors: emotional attachment, conditioning, epistemic curiosity, intrinsic innovation orientation, intrinsic technology acceptance, and empathy.

SVM achieved slightly lower but comparable performance (Table 1). Mean accuracy was 81% (SD = 0.9%, range: 80–84%) and mean AUC was 91% (SD = 0.5%, range: 90–92%). Variable importance rankings differed slightly, identifying emotional attachment, conditioning, epistemic curiosity, empathy, intrinsic innovation orientation, and optimism orientation as predictive features.

MARS exhibited the most variability and lowest performance (Table 1). Mean accuracy was 80% (SD = 1.53%, range: 76–86%) with mean AUC of 88% (SD = 1.31%, range: 85–90%). This model identified only two important variables: emotional attachment and intrinsic innovation orientation.

RF achieved the highest mean accuracy (83%) with lowest variance (0.8%), indicating both superior performance and stability across random samples. Variable importance analysis identified functional value, emotional value, and consumer innovativeness as dominant predictors across algorithms.

Table 1. Classification Model Performance Comparison.

Model	Metric	Worst	Best	Mean (sd)
RF	Accuracy	81%	85%	83% (0.8%)
	AUC	90%	93%	91% (0.5%)
SVM	Accuracy	80%	84%	81% (0.9%)
	AUC	90%	92%	91% (0.5%)
MARS	Accuracy	76%	86%	80% (1.53%)
	AUC	85%	90%	88% (1.31%)

4.2 Reduced-Feature Model: The Parsimonious Solution

If a decision maker can identify which questions are sufficient to estimate the probability that a consumer will enter (or not enter) a gamification process, those questions should form a small yet informative set. Because evaluating consumers' willingness to engage in a gamification process may involve survey-based measurement, it is important to rely on a minimal but meaningful set of variables in order to reduce respondent burden and encourage participation. Accordingly, RF were tested to assess whether they could maintain or improve performance using only the three most important variables: functional value, emotional value, and consumer innovativeness. Furthermore, reducing the number of input variables in a RF classification problem can enhance predictive performance by minimizing noise, reducing overfitting, and improving generalization. Irrelevant or redundant features introduce unnecessary randomness into the tree-building process, weakening split quality and ultimately degrading model performance. By concentrating on the most informative variables, the model avoids learning spurious associations, leading to greater stability and reduced variance across trees. Moreover, removing features that are not contributing to the final outcome helps to diminish the curse of dimensionality, thereby preserving well-defined decision boundaries and improving computational efficiency. Empirical validation—such as comparing classification accuracy before and after feature reduction—strongly supports this approach. The obtained results show that feature reduction not only maintained performance but improved accuracy, as can be seen in Table 2.

Table 2. Reduced-Feature Model Performance.

Model	Metric	Worst	Best	Mean (sd)
RF	Accuracy	82%	87%	84% (0.88%)
	AUC	89%	92%	91% (0.47%)

Statistical significance testing confirmed these improvements were indeed relevant: Wilcoxon signed-rank test (Accuracy): $p = 3.014 \times 10^{-11}$; Paired t-test (AUC): $p =$

6.021×10^{-16}. Both differences are highly statistically significant ($p < 0.001$), indicating that the three-variable model generalizes better than the full-feature model. This improvement aligns with machine learning theory: eliminating irrelevant or redundant features reduces noise, minimizes overfitting, and decreases variance across decision trees, enabling the algorithm to identify more robust decision boundaries and avoid spurious correlations.

Because we are able to construct a predictive model that relies on only three variables while maintaining reliable performance, firms can assess consumers' willingness to enter a gamification process by asking a very limited number of questions, specifically those measuring functional value (*"It will be useful to me"*, *"It will be easy to use"*, *"It will give me advantages"*), emotional value (*"It will excite me"*, *"It will provide me with varied enjoyment"*, *"It will be fun"*) and consumer innovativeness (*"Other people come to me for advice on new technologies"*, *"In general, I am among the first in my circle of friends to acquire new technology when it appears"*, *"I can usually figure out new high-tech products and services without help from others"*, *"I keep up with the latest technological developments in my areas of interest"*).

4.3 Decision Tree Interpretation

To interpret the predictive patterns of the reduced model, a decision tree was trained using the same subset of attributes previously selected for the RF model. While the RF served as the primary predictor due to its superior performance, the decision tree offers an interpretable representation of how the model differentiates between classes. Random forests perform very well considering their predictive performance because they combine hundreds of decision trees, each trained on a bootstrap sample with a random subset of features. This ensemble approach increases robustness to noise, allows random forests to be able to represent complex nonlinear relationships, and allows evaluation of feature importance. However, interpretability of the results is rather limited, since it is difficult to analyze and interpret the model as a whole. A single decision tree, by contrast, offers explicit if-then rules that can be visualized and communicated, although most of the time it is not able to reach the same level of classification accuracy obtained with RF. One possible compromise solution is to use RF to identify the most relevant attributes and then training a constrained decision tree on that reduced set. This strategy is especially valuable when stakeholders, like managers, prefer rule-based explanations and transparent decision logic. This is exactly what was done.

The resulting tree (Fig. 1) reveals the hierarchical relationships among key psychological and behavioral variables. The first split, based on the functionality score (*Func*), separates participants into two primary groups, showing that higher perceived functionality is strongly associated with the positive class. Subsequent splits refine the classification according to emotional engagement (*emot*) and innovation-related traits (*innovtr*). Participants with low functionality and emotional engagement are mostly classified in the negative class, while those with higher functionality or innovation scores are more likely to belong to the positive class. Actually, is it known that consumer innovativeness - an inherent personal characteristic reflecting one's willingness to pursue, accept, and adopt new products, services, ideas, or behaviors independently of peer influence – is able to consistently predict adoption of market innovations [19]. Prior research shows that

highly innovative consumers are more receptive to novel marketing approaches, including gamification. This personality trait complements consumption value dimensions by capturing autonomous decision-making regarding innovation adoption.

Terminal nodes in the tree display the number of observations per class and the proportion of correctly classified cases, confirming that the positive class is more prevalent among participants with higher scores on these dimensions. Overall, the decision tree represents functionality, emotional engagement, and innovation as very important factors for distinguishing participants likely to engage (positive class) from those less likely (negative class). Functional value serves as the primary decision node: participants with low perceived functionality are generally classified as non-engaging, whereas those with high functionality are further evaluated on emotional engagement. High scores on both functionality and emotional engagement correspond to a very high probability of gamification acceptance. Consumer innovativeness plays an important role in reaching a classification in lower levels of the decision tree.

This hierarchical structure provides managers with an intuitive and transparent assessment tool. By querying participants on these three dimensions through a brief questionnaire, organizations can make rapid, confident predictions regarding gamification suitability.

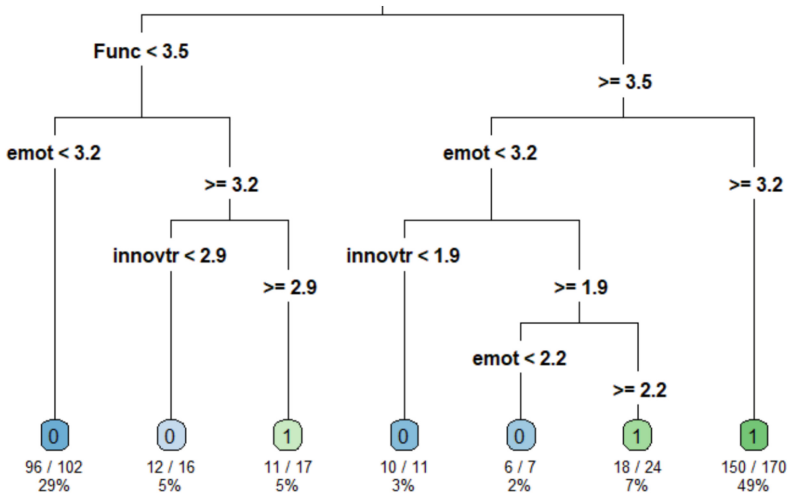


Fig. 1. Decision tree classifying participants based on functionality, emotional engagement, and innovativeness, with terminal nodes showing class proportions

5 Managerial Implications

This research offers several practical insights for marketing practitioners:

- **Pre-Implementation Market Assessment:** Before investing resources in gamification development, managers can use a brief survey to assess functional value, emotional value, and consumer innovativeness. Based on these three dimensions, our model predicts whether individual clients are likely to adopt gamification with an

accuracy of 84%, supporting more informed go/no-go decisions and helping identify segments with lower adoption potential.

- **Market Segmentation and Targeting:** Rather than implementing broad, organization-wide gamification strategies, managers should identify customer segments exhibiting high functional value perception, high emotional value orientation, and high consumer innovativeness. This targeted approach concentrates resources on customers with a high-probability of engaging in the gamification strategy, maximizing return on investment.
- **Gamification Design Strategy:** For identified high-probability segments, gamification design should emphasize both functional benefits (efficient reward accumulation, clear progression paths, ease of use) and emotional appeal (entertainment, enjoyment, fun narratives). On the other hand, for lower-probability segments, alternative engagement strategies may prove more cost-effective than gamification strategies.
- **Risk Mitigation:** Implementing gamification prematurely and without previously studying the potential impact on customers may harm the brand, namely if customers perceive its mechanics as intrusive or poorly aligned with their preferences. By segmenting audiences according to their demonstrated receptivity, organizations can reduce this reputational risk while improving campaign effectiveness.

6 Discussion

The work developed presents interesting results considering the adoption of gamification within a marketing context, since it allows organizations to base their decisions regarding the development of gamification strategies grounded on data-driven models able to robustly predict the potential costumer engagement. Although the questionnaire data could have been analyzed using a classical structural equation modeling approach, this work adopts a practical perspective, using the questionnaire as an operational tool to identify consumers' willingness to engage in a gamification process. Most of existing research works document gamification's benefits for engaged consumers, whilst the current work is focused on providing guidance for ex-ante evaluation of adoption likelihood within specific markets.

The theoretical contribution lies in demonstrating that consumption value theory - functional and emotional dimensions - combined with consumer innovativeness provides a simple yet powerful framework for predicting gamification adoption. These constructs capture the existing heterogeneity in how consumers evaluate novel marketing approaches, explaining the differences that can be observed in gamification effectiveness. This work contributes methodologically by showing that, with just a very small set of carefully selected questions, it is possible to achieve highly accurate predictions of consumer behavior. By reducing features to only three key variables, the RF model attains superior classification performance - higher AUC and lower variance - compared to the full six-variable model, highlighting both the practical efficiency and predictive power of the method. The practical contribution is a lightweight, implementable tool: a small questionnaire that feeds a trained RF model able to output an accurate assessment of whether gamification aligns with individual or segment-level customer preferences. This represents a substantial advance over existing practice, where such decisions typically rely on intuition rather than evidence.

Although the results presented are interesting and useful, it is also important to acknowledge some existing limitations. Firstly, the sample (N = 347) is moderate in size; validation on larger, more diverse contexts and populations would strengthen generalizability. Secondly, this study measures *stated* willingness to engage via questionnaire; actual behavioral response to deployed gamification may differ due to implementation quality and contextual factors. Thirdly, while we tested three classes of algorithms, exploration of deep learning architectures might reveal further performance improvements, although requiring larger datasets. Fourthly, this cross-sectional analysis does not capture temporal dynamics: consumer preferences and technology adoption propensity evolve over time. Future research should examine whether the three-variable model generalizes across different contexts and demographic populations, validate its predictive power in real-world settings by comparing predicted and actual gamification adoption behavior, integrate external variables such as the competitive landscape, media exposure, and social influence into the predictive framework, and conduct longitudinal studies to analyze how innovativeness and value perceptions evolve alongside technological change.

7 Conclusion

This research demonstrates that machine learning classification models can reliably predict consumer willingness to adopt gamification using a parsimonious set of theoretically grounded variables: functional value, emotional value, and consumer innovativeness. RF achieved 84% prediction accuracy with an AUC of 0.91, clearly better than random guessing (50% on balanced data) and providing actionable guidance for managers. The practical payoff is significant: organizations can conduct rapid, evidence-based market assessment before spending substantial resources with gamification development. By segmenting audiences and targeting gamification to high-probability adopters, companies enhance return on investment, minimize brand misalignment risk, and concentrate efforts on customer segments with a higher probability of engaging. Machine learning predictions can be converted into decision rules that managers can more easily understand, making them more useful in practice.

As the adoption of marketing gamification strategies continues to grow, especially across digital markets, data-driven approaches capable of predicting customer adoption become increasingly valuable. This work provides both theoretical understanding of customer adoption or rejection of gamification strategies as well as an example of a practical tool that can be used for market assessment, enabling organizations to make informed, risk-mitigated decisions regarding this powerful yet costly marketing innovation.

Acknowledgments. This work has been funded by national funds through FCT – Fundação para a CiênciaeaTecnologia,I.P.,throughprojectUIDB/05037/2025withDOI10.54499/UIDB/05037/2025 and UID/00308/2025 INESC Coimbra, DOI 10.54499/UIDB/00308/2025.

Disclosure of Interests. The authors have no competing interests.

Appendix 1

STUDY ABOUT ONLINE SHOPPING BEHAVIOR

I. The following statements refer to your perception of the company.

Please indicate your degree of agreement with the following statements:		Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree
1	I trust this company.	1	2	3	4	5
2	I rely on this company.	1	2	3	4	5
3	This is an honest company.	1	2	3	4	5
4	This company is safe.	1	2	3	4	5

II. Please read the following scenario carefully:

Suppose that the company's website/application provides you an interactive, enjoyable and totally free gaming experience. By creating a user profile, you become part of a community that allows you to interact with other users, exchange information and recommendations and add friends from your social networks. You will also be able to participate in the improvement and creation of products/services and in the development of solutions offered by this organization. Over the course of time, you will be invited to fulfill various challenges which, in addition to improving your shopping experience, making it even more rewarding, will give you points. These challenges include, for example: (1) sharing the website link with your family and friends; (2) evaluating and commenting on the products and/or services you have purchased; (3) reaching a certain amount in purchases per month, among others.

The more points you gain, the more benefits you will have, achieving new heights as a customer. Through your involvement, you will have access to exclusive promotions, priority access to new products, surprise gifts, among others. You can also obtain badges that symbolize your achievements and challenge your friends to a competition or to cooperate towards a certain objective. On your profile you will be able to track your progress and compare it with that of other users. Moreover, you will be able to share your achievements on social media.

III. Assuming that the previously described gaming experience is currently available to you. Please indicate your degree of agreement with the following statements:

It seems to me that integrating this gaming experience:		Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree
1	It will be useful to me.	1	2	3	4	5
2	It will be easy to use.	1	2	3	4	5
3	It will give me advantages.	1	2	3	4	5
4	It will excite me.	1	2	3	4	5
5	It will provide me with varied enjoyment.	1	2	3	4	5
6	It will be fun.	1	2	3	4	5
7	It will help me to feel acceptable.	1	2	3	4	5
8	It will improve the way that I am perceived.	1	2	3	4	5
9	It will make a good impression on other people.	1	2	3	4	5
10	It will give social approval.	1	2	3	4	5
11	It will arouse my curiosity for novelty.	1	2	3	4	5
12	It will provide me with a new experience.	1	2	3	4	5
13	It will help me acquire new knowledge.	1	2	3	4	5
14	It will make my purchasing process easier.	1	2	3	4	5
15	It will help me to find information easily.	1	2	3	4	5
16	It will help me to achieve information quickly.	1	2	3	4	5

IV. Continue to assume that the gaming experience described above is currently available to you. Please indicate your degree of agreement with the following statements:

		Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree
1	I plan to enter in this experience in the future.	1	2	3	4	5
2	I intend to enter in this experience in the future.	1	2	3	4	5
3	I predict I will enter in this experience in the future.	1	2	3	4	5

V. Next, you will find a list of statements about how you perceive TECHNOLOGY.

	Indicate your level of agreement with each of the statements:	Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree
1	New technologies contribute to a better quality of life.	1	2	3	4	5
2	Technology gives me more freedom of mobility.	1	2	3	4	5
3	Technology gives people more control over their daily lives.	1	2	3	4	5
4	Technology makes me more productive in my personal life.	1	2	3	4	5
5	Other people come to me for advice on new technologies.	1	2	3	4	5
6	In general, I am among the first in my circle of friends to acquire new technology when it appears.	1	2	3	4	5
7	I can usually figure out new high-tech products and services without help from others.	1	2	3	4	5
8	I keep up with the latest technological developments in my areas of interest.	1	2	3	4	5
9	People are too dependent on technology to do things for them.	1	2	3	4	5
10	Too much technology distracts people to a point that is harmful.	1	2	3	4	5
11	Technology lowers the quality of relationships by reducing personal interaction.	1	2	3	4	5
12	I do not feel confident doing business with a place that can only be reached online.	1	2	3	4	5
13	When I get technical support from a provider of a high-tech product or service, I sometimes feel as if I am being taken advantage of by someone who knows more than I do.	1	2	3	4	5
14	Technical support lines are not helpful because they don't explain things in terms I understand.	1	2	3	4	5
15	Sometimes, I think that technology systems are not designed for use by ordinary people.	1	2	3	4	5
16	There is no such thing as a manual for a high-tech product or service that's written in plain language.	1	2	3	4	5

VI. The following statements address your relationship with VIDEO GAMES.

	Indicate your level of agreement with each of the statements:	Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree
1	I like playing video games.	1	2	3	4	5
2	I often play video games.	1	2	3	4	5
3	Compared to people of my age, I spend more time playing video games.	1	2	3	4	5
4	I would describe myself as a gamer.	1	2	3	4	5
5	I play different types of video games.	1	2	3	4	5

VII. Now consider your way of being. Please indicate the degree to which you generally feel as follows:

		Very little or nothing	A little	Moderately	Quite	Extremely
1	Inspired	1	2	3	4	5
2	Attentive	1	2	3	4	5
3	Excited	1	2	3	4	5
4	Enthusiastic	1	2	3	4	5
5	Determined	1	2	3	4	5
6	Afraid	1	2	3	4	5
7	Upset	1	2	3	4	5
8	Nervous	1	2	3	4	5
9	Scared	1	2	3	4	5
10	Distressed	1	2	3	4	5

VIII. Think again about the way you are.

Indicate your level of agreement with each of the statements:		Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree
1	I am a more melancholic person than others.	1	2	3	4	5
2	I am a temperamental person.	1	2	3	4	5
3	I am an envious person.	1	2	3	4	5
4	My emotions go way up and down.	1	2	3	4	5
5	I am more irritable than other people.	1	2	3	4	5
6	I am less shy person than others.	1	2	3	4	5
7	I am a daring person.	1	2	3	4	5
8	I am an extroverted person.	1	2	3	4	5
9	I feel I am a very creative person.	1	2	3	4	5
10	I always have new ideas.	1	2	3	4	5
11	I appreciate art.	1	2	3	4	5
12	I find new solutions to problems.	1	2	3	4	5
13	I am more original than other people.	1	2	3	4	5
14	I am kind to other people.	1	2	3	4	5
15	I am tenderhearted with others	1	2	3	4	5
16	I am sympathetic.	1	2	3	4	5
17	I am a precise person.	1	2	3	4	5
18	I am an organized person.	1	2	3	4	5
19	I am a sloppy person.	1	2	3	4	5
20	I am an orderly person.	1	2	3	4	5

Indicate your level of agreement with each of the statements:		Totally disagree	Disagree	Neither agree nor disagree	Agree	Totally agree
1	I don't like to make decisions quickly.	1	2	3	4	5
2	I don't like to take part in really fast-paced conversations.	1	2	3	4	5
3	I don't like to do things quickly.	1	2	3	4	5
4	I try to avoid activities where I have to act without much time to think first.	1	2	3	4	5
5	I frequently make appointments without thinking about whether I will be able to keep them.	1	2	3	4	5
6	I often make up my mind without taking the time to consider the situation from all angles.	1	2	3	4	5
7	Many times the plans I make don't work out because I haven't gone over them carefully enough in advance.	1	2	3	4	5
8	I often get involved in projects without first considering the potential problems.	1	2	3	4	5

IX. Finally, we ask you for some information so that we can characterize the profile of the respondents.

1. Gender:
☐ Female
☐ Male
☐ Other

2. Age: _____ years

3. Nationality:
_____

4. Number of people in the household:
☐ 1 ☐ 2 ☐ 3
☐ 4 ☐ 5 ☐ 6 or more

5. Profession:
☐ Student
☐ Worker/student

6. What level of education do you attend?
☐ Degree
☐ Master's degree
☐ PhD
☐ Other. Which? _____

7. Please tell us the name of the course and year you are attending:
_____
_____

8. Net monthly income of the household:
☐ Less than €500
☐ €500-999
☐ €1000-1499
☐ €1500-2499
☐ €2500-4999
☐ €5000 or more

9. In the last 12 months, how many times have you purchased products/services online?
☐ Never
☐ 1 to 3 times
☐ 4 to 6 times
☐ 7 to 9 times
☐ 10 to 12 times
☐ 13 or more times

Thank you for your cooperation.

References

1. Deterding, S., Dixon, D., Khaled, R., Nacke, L.: From game design elements to gamefulness: defining "gamification." In: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments. ACM, Tampere, Finland, pp. 9–15 (2011)
2. Gupta, A.S., Mukherjee, J.: Long-term changes in consumers' shopping behavior post-pandemic: an exploratory study. Int. J. Retail Distrib. Manag. **50**, 1518–1534 (2022)
3. Mordor Intelligence: Gamification Market Size & Share Analysis - Growth Trends and Forecast (2026–2031)
4. Hsu, C.-L., Chen, M.-C.: How does gamification improve user experience? An empirical investigation on the antecedences and consequences of user experience and its mediating role. Technol. Forecast. Soc. Chang. **132**, 118–129 (2018)
5. Xi, N., Hamari, J.: Does gamification affect brand engagement and equity? A study in online brand communities. J. Bus. Res. **109**, 449–460 (2020)
6. Rialti, R., Filieri, R., Zollo, L., et al.: Assessing the relationship between gamified advertising and in-app purchases: a consumers' benefits-based perspective. Int. J. Advert. **41**, 868–891 (2022)
7. Behl, A., Pereira, V., Jayawardena, N., et al.: Gamification as an innovation: a tool to improve organizational marketing performance and sustainability of international firms. Int. Mark. Rev. **41**, 107–137 (2024)
8. Witoszek-Kubicka, A.: Implementation of gamification in polish companies—stages, elements. Ethics. Inf. **11**, 371 (2020)
9. Vilkaite-Vaitone, N., Kirse, S., Adomaviciute-Sakalauske, K., et al.: The usefulness of gamification for enhancing customer loyalty to small e-tailers. EuroMed J. Bus. **20**, 435–452 (2025)
10. Zhang, L., Shao, Z., Chen, B.: Exploring the role of immersion-based gamification in enhancing online consumer perceived value and behavioral intentions: a task-technology fit perspective. Int. J. Hum.–Comput. Interact., 1–17 (2025)
11. Bitrián, P., Buil, I., Catalán, S.: Enhancing user engagement: the role of gamification in mobile apps. J. Bus. Res. **132**, 170–185 (2021)
12. Jang, S., Kitchen, P.J., Kim, J.: The effects of gamified customer benefits and characteristics on behavioral engagement and purchase: evidence from mobile exercise application uses. J. Bus. Res. **92**, 250–259 (2018)
13. Srivastava, G., Bag, S.: Diagnosing key factors for gamification in marketing using hierarchical clustering technique. IJTMKT **15**, 354–378 (2021)
14. Ryan, R.M., Deci, E.L.: Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp. Educ. Psychol. **25**, 54–67 (2000)
15. Yang, Y., Asaad, Y., Dwivedi, Y.: Examining the impact of gamification on intention of engagement and brand attitude in the marketing context. Comput. Hum. Behav. **73**, 459–469 (2017)
16. Sheth, J.N., Newman, B.I., Gross, B.L.: Why we buy what we buy: a theory of consumption values. J. Bus. Res. **22**, 159–170 (1991)
17. Wang, H.-Y., Liao, C., Yang, L.-H.: What affects mobile application use? The roles of consumption values. Int. J. Market. Stud. **5**, 11 (2013)
18. Koay, K.Y., Leong, M.K.: Understanding consumers' intentions to use drone food delivery services: a perspective of the theory of consumption values. Asia-Pac. Jo. Bus. Adm. **16**, 1226–1240 (2024)
19. Kaur, P., Dhir, A., Talwar, S., Ghuman, K.: The value proposition of food delivery apps from the perspective of theory of consumption value. IJCHM **33**, 1129–1159 (2021)

20. Omigie, N.O., Zo, H., Rho, J.J., Ciganek, A.P.: Customer pre-adoption choice behavior for M-PESA mobile financial services: extending the theory of consumption values. Ind. Manag. Data Syst. **117**, 910–926 (2017)
21. Thongmak, M.: Determinants of intention to play Pokémon Go. Heliyon **6**, 12 (2020)
22. Kamboj, S., Rana, S., Drave, V.A.: Factors driving consumer engagement and intentions with gamification of mobile apps. J. Electron. Commer. Organ. (JECO) **18**, 17–35 (2020)
23. Chaudhuri, A., Holbrook, M.B.: The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. J. Mark. **65**, 81–93 (2001)
24. Mackinnon, A., Jorm, A.F., Christensen, H., et al.: A short form of the positive and negative affect schedule: evaluation of factorial validity and invariance across demographic variables in a community sample. Pers. Individ. Differ. **27**, 405–416 (1999)
25. Hur, W.-M., Yoo, J.-J., Chung, T.-L.: The consumption values and consumer innovativeness on convergence products. Ind. Manag. Data Syst. **112**, 688–706 (2012)
26. Lin, P.-C., Huang, Y.-H.: The influence factors on choice behavior regarding green products based on the theory of consumption values. J. Clean. Prod. **22**, 11–18 (2012)
27. Parasuraman, A., Colby, C.L.: An updated and streamlined technology readiness index: TRI 2.0. J. Serv. Res. **18**, 59–74 (2015)
28. Kuem, J., Ray, S.: Personality antecedents and consequences of Internet addiction vis-à-vis internet habit: a theory and an empirical investigation. Inf. Syst. Front. **24**, 579–594 (2022)
29. Mowen, J.C., Spears, N.: Understanding compulsive buying among college students: a hierarchical approach. J. Consum. Psychol. **8**, 407–430 (1999)
30. Bourgonjon, J., Valcke, M., Soetaert, R., Schellens, T.: Students' perceptions about the use of video games in the classroom. Comput. Educ. **54**, 1145–1156 (2010)
31. Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: User acceptance of computer technology: a comparison of two theoretical models. Manage. Sci. **35**, 982–1003 (1989)
32. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
33. Vapnik, V.: The Nature of Statistical Learning Theory. Springer Science & Business Media (2013)
34. Friedman, J.H.: Multivariate adaptive regression splines. Ann. Stat. **19**, 1–67 (1991)

Multi-objective Optimization of Distributed Job Shop Scheduling Problems with Customer Time Requirements

Francisco dos Santos[1,2(✉)], Lino Costa[1,3], and Leonilde Varela[1,3]

[1] ALGORITMI Research Centre/LASI, University of Minho, Braga, Portugal
{lac,leonilde}@dps.uminho.pt
[2] Polytechnic Institute, Kimpa Vita University, Uíge, Angola
francisco_dos_santos@outlook.pt
[3] Department of Production and Systems, University of Minho, Braga, Portugal

Abstract. In the contemporary industrial context, characterized by intense competitive pressures, consumers have become increasingly discerning, imposing stringent criteria on suppliers regarding delivery times, quality standards, and costs. In recent years, in response to this growing demand and the search for greater operational efficiency, several organizations have chosen to fragment their production units, migrating to decentralized manufacturing models. In this context, production systems are characterized by the need to coordinate multiple operations that are processed on machines distributed across different manufacturing locations. The present paper proposes a multi-objective model for the job shop scheduling problem in a distributed context, incorporating uncorrelated parallel machines and setup times influenced by jobs ordering. The model incorporates a temporal dimension that accounts for the time required to transport raw materials to the manufacturing unit selected for each operation, as well as the deadlines set by customers. The integration of customer priorities, materialized in the simultaneous optimization of three indicators: total processing time (makespan), delivery delays (tardiness) and production advances (earliness), represents a novel innovation of this study. In order to validate the proposed methodology, NSGA-III was used to solve small instances of the problem. The preliminary findings substantiate the efficacy of the proposed strategy.

Keywords: Distributed job shop scheduling problem · multi-objective optimization · evolutionary algorithms

This work has been supported by FCT Fundação para a Ciência e Tecnologia within the R&D Unit Project Scope UID/00319/2025 - Centro ALGORITMI (ALGORITMI/UM) https://doi.org/10.54499/UID/00319/2025..

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 182–198, 2027.
https://doi.org/10.1007/978-3-032-30521-3_12

1 Introduction

In recent decades, the industrial sector has undergone significant transformations driven by market globalization and technological advances. This new economic paradigm is characterized by fierce competition, where organizations are constantly challenged to offer high-quality products, short delivery times, and competitive costs thanks to the use of job shop scheduling in production processes [13]. In this dynamic environment, adaptability and operational efficiency have become critical factors for survival and success in these industries [18].

At the same time, a structural change in the organization of production has been observed. Traditionally concentrated in single manufacturing units, which are generally traditional job shop scheduling problems (JSS) [15], manufacturing has evolved toward decentralized models, where companies distribute their operations across multiple geographic locations. This strategy, known as distributed production, not only increases production capacity, but also brings production closer to consumer markets, reduces logistics costs, and mitigates risks associated with localized contingencies. However, this decentralization introduces additional complexities in production planning and control, particularly with regard to the coordination of jobs between different manufacturing units.

In this context, the problem of distributed job shop scheduling problem (DJSS) arises [4,7,8], involving different optimization criteria such as costs, delivery time, delay, among others [6,17]. Unlike traditional job shop scheduling problems [10], where all machines are located in a single location, DJSS problems involve the allocation and sequencing of jobs on machines distributed across different factories [16], adding an additional layer of complexity related to the choice of production unit for each operation and the transport times involved [12]. The existing literature has devoted considerable attention to the optimization of classic performance indicators such as makespan (total completion time). However, in a customer-oriented market, other criteria are becoming increasingly relevant. Meeting delivery deadlines has become a key competitive advantage, involving not only minimizing delays (tardiness) but also reducing production advances (earliness). Early deliveries can generate storage and capital immobilization costs, while late deliveries compromise customer satisfaction and the company's reputation. The simultaneous consideration of these factors, aligned with customer preferences, represents an important gap in current research.

Incorporating customer preferences into this environment requires the simultaneous optimization of multiple objectives. In addition to minimizing makespan, it is essential to minimize tardiness, ensuring that deliveries occur within promised deadlines and earliness, avoiding the storage and capital immobilization costs associated with early production. The joint consideration of these three criteria, in a context of unrelated parallel machines, sequence-dependent setup times, and transport times, presents a highly complex problem, the solution of which requires sophisticated multi-objective approaches, such as Non-dominated Sorting Genetic Algorithm III (NSGA-III) [5,15].

This paper proposes a multi-objective model for the distributed job shop scheduling problem that incorporates unrelated parallel machines, sequence-

dependent setup times [1], raw material transport times, and customer order delivery times. As a main, the model integrates customer preferences such as meeting delivery deadlines and reducing warehousing costs through the simultaneous optimization of three objectives: minimization of makespan, minimization of tardiness, and minimization of earliness. To solve the problem, NSGA-III is used, a metaheuristic recognized for its effectiveness in addressing complex multi-objective problems.

This paper is organized as follows: in addition to the introduction, Sect. 2 presents a multi-objective optimization model, including an overview of the relevant literature on distributed job shop scheduling problems, as well as a mathematical formulation of the problem, and the optimization approach. Section 3 reports and analyzes the experimental results for three instances of the problem. Finally, Sect. 4 addresses the conclusions and suggestions for future work.

2 Multi-objective Optimization Model

2.1 Distributed Job Shop Scheduling

The traditional job shop scheduling problem is one of the most studied classes in the field of production optimization, recognized for its computational complexity and practical relevance. In its classic formulation, JSS problems involve a set of jobs that must be processed on a given number of machines, following a predefined sequence for each job [3,7,10]. Each operation requires a specific machine for a given period of time, and the central objective is to find the processing sequence that optimizes certain performance criteria. Traditionally, the most commonly used criterion has been the minimization of makespan, i.e., the total time required to complete all tasks. However, the evolution of production systems and market demands has broadened this spectrum, incorporating metrics related to meeting deadlines, energy efficiency, or load balancing between resources [14].

The transition to decentralized production models gave rise to the distributed job shop scheduling problem, an extension of JSS problems that reflects contemporary industrial reality. In this scenario, machines are not all concentrated in a single location, but rather distributed across multiple factories, geographically dispersed [16,17]. This configuration introduces two interdependent decisions: the assignment of each task to a manufacturing unit and the sequencing of operations within each factory. The complexity of the problem is further amplified by the need to consider transport times between different locations, as well as the possibility that each factory may have multiple parallel machines capable of performing the same operation with potentially different processing times (unrelated parallel machines). Additionally, the setup times of the machines can vary depending on the sequence of tasks processed, reflecting the need for adjustments, tool changes, or cleaning between consecutive operations [1,10].

In addition to the structural complexities of DJSS problems, customer preferences emerge as a determining factor in the definition of optimization objectives. In a service-oriented market, strict compliance with delivery deadlines has

become strategic importance. This reality translates into the need to minimize not only tardiness, which represents deliveries made after the date promised to the customer, but also earliness, which corresponds to early completion of production [17]. Although early delivery may seem benign at first glance, it entails significant costs related to storage, additional handling, and capital immobilization. The simultaneous optimization of makespan, tardiness, and earliness poses a complex multi-objective problem, where there are often trade-offs between the different criteria: for example, minimizing makespan can lead to solutions that compromise the fulfillment of individual deadlines. The integrated consideration of these objectives, aligned with customer priorities, represents a significant advance over traditional approaches focused exclusively on internal production efficiency.

2.2 Mathematical Formulation

In traditional job shop scheduling problems, there is a set of jobs that must be executed on a set of machines, following predefined processing sequences. The classic objective generally focuses on minimizing the total completion time of all jobs, known as the makespan. However, this traditional approach does not adequately address the requirements associated with meeting delivery deadlines, ignoring both the costs resulting from delays (tardiness) and the inefficiencies caused by early completion (earliness).

In distributed job shop scheduling problems, complexity increases significantly. In addition to the set of jobs that must be executed on a set of machines, these machines are dispersed across multiple factories located in different geographic regions. This dispersion introduces additional decisions related to the assignment of each job to a specific manufacturing unit. Consequently, in addition to the processing and setup times of the jobs on the machines, it is imperative to consider other factors, such as the transport times of raw materials between factories or to production units, as well as the implications of this geographic distribution on meeting deadlines agreed upon with customers. Thus, a DJSS problem involves the following sets:

- $\mathbb{N} = \{n_1, n_2, n_3, ...n_N\}$ with N jobs;
- $\mathbb{F} = \{f_1, f_2, f_3, ...f_F\}$ with F factories;
- $\mathbb{M}_1 = \{m_{1,1}, \ldots, m_{1,M_1}\}$, $\mathbb{M}_2 = \{m_{2,1}, \ldots, m_{2,M_2}\}$, ..., $\mathbb{M}_F = \{m_{F,1}, \ldots, m_{F,M_F}\}$ with $M_1, M_2, \ldots, M_F$ machines for each factory, respectively.

It is assumed that each machine in each factory is continuously available and can execute only one job at a time. Moreover, all jobs are independent, and none of them have precedence constraints. Preemption is not allowed; that is, the processing of any job on a machine cannot be interrupted. It also assumed that the following problem data is known:

- $T_{F \times N}$ matrix, where $t_{u,j}$ is the transportation time to transport job j to the factory u;

- $P_{F\times M\times N}$ matrix, where $p_{u,i,j}$ is the processing time of job j on machine i of factory u;
- $S_{F\times M\times N\times N}$ matrix, where $s_{u,i,j,k}$ is the setup time to process job j after job k on machine i of factory u ($s_{u,i,j,0}$ is the setup time to process job j at the initial instant on machine i of factory u);
- D_N vector, where d_j is the delivery date of the job j.

The goal is to find a feasible sequence of jobs on each machine in different factories that minimizes the makespan [12,17], the tardiness time and the earliness [17]. The feasible sequence π assigns all jobs to machines in each factory, i.e., it is the vector of all feasible subsequences $\pi_{u,i}$ in each machine i and factory u. For a given sequence π, the makespan ($c_{\max}(\pi)$) can be expressed as follows:

$$c_{\max}(\pi) = \max_{u=1,\ldots,F} \Big(\max_{i=1,\ldots,M_u} (c_{u,i}(\pi)) \Big) \tag{1}$$

where $c_{u,i}(\pi)$ is the total completion time of the jobs on machine i of the factory u for the feasible sequence π. This value can be computed by

$$c_{u,i}(\pi) = \sum_{j\in\pi_{u,i}} t_{u,j} + \sum_{j\in\pi_{u,i}} p_{u,i,j} + \sum_{j\in\pi_{u,i}} s_{u,i,j,k}$$

with k being the jobs that precede jobs j in the subsequence $\pi_{u,i}$ of the feasible sequence π. It should be noted that when the job j is the first job in the subsequence $\pi_{u,i}$, k is zero, and therefore $s_{u,i,j,0}$ is the initial setup time.

The tardiness time ($t_{\max}(\pi)$) for a given sequence π is given by

$$t_{\max}(\pi) = \max_{u=1,\ldots,F} \Big(\max_{i=1,\ldots,M_u} (t_{u,i}(\pi)) \Big) \tag{2}$$

where $t_{u,i}(\pi) = \max_{j\in\pi_{u,i}} (0, c^j_{u,i}(\pi) - d_j)$, and

$$c^j_{u,i}(\pi) = \sum_{\forall z\preceq j\wedge z,j\in\pi_{u,i}} t_{u,j} + \sum_{\forall z\preceq j\wedge z,j\in\pi_{u,i}} p_{u,i,z} + \sum_{\forall z\preceq j\wedge z,j\in\pi_{u,i}} s_{u,i,z,k}$$

is the completion time of the job j, with z being all jobs that precede or are equal to the job j in the subsequence $\pi_{u,i}$ of the feasible sequence π.

The earliness time ($l_{\max}(\pi)$) for a given sequence π is defined by

$$l_{\max}(\pi) = \max_{u=1,\ldots,F} \Big(\max_{i=1,\ldots,M_u} (l_{u,i}(\pi)) \Big) \tag{3}$$

where $l_{u,i}(\pi) = \max_{j\in\pi_{u,i}} (0, d_j - c^j_{u,i}(\pi))$.

In this paper, two bi-objective DJSS problems are considered:

$$\min \mathbf{F_1}(\pi) = \big(c_{\max}(\pi), t_{\max}(\pi)\big)^T, \tag{4}$$

and

$$\min \mathbf{F_2}(\pi) = \big(c_{\max}(\pi), l_{\max}(\pi)\big)^T. \tag{5}$$

Additionally, the following multi-objective DJSS problem with three objectives is also addressed:

$$\min \mathbf{F_3}(\pi) = \big(c_{\max}(\pi), t_{\max}(\pi), l_{\max}(\pi)\big)^T. \tag{6}$$

In these formulations, π is a feasible sequence that belongs to the feasible set Ω (the set of all possible permutations) and $\mathbf{F_1}(\pi)$, $\mathbf{F_2}(\pi)$, and $\mathbf{F_3}(\pi)$ are the vectors of objective functions defined in the objective space.

2.3 Optimization Approach

DJSS problems with customer time requirements involve several conflicting objectives such as makespan, tardiness and earliness expressed by equations (1), (2) and (3). Multi-objective optimization algorithms aim to find optimal trade-offs between objectives, the so-called Pareto-optimal set. The images of the Pareto-optimal solutions define a Pareto optimal frontier in the objective space. This frontier allows the decision-maker to select a solution taking into account the several optimal alternatives that best fulfill his preferences [9].

Evolutionary algorithms draw inspiration from the genetic mechanisms of living organisms, which evolve over time through natural selection. In particular, genetic algorithms are widely applied to multi-objective problems because their working principles are based on the evolution of a population of chromosomes representing candidate solutions and the application of genetic operators to generate new ones. Therefore, it is possible to approximate the Pareto-optimal set effectively in a single run.

Furthermore, DJSS problems are also highly complex combinatorial problems. Certain features of evolutionary algorithms make them particularly suitable for addressing such problems, since constraints can be managed through permutation-based chromosome representations, and genetic operators can be specifically tailored to these search spaces. This enables the algorithm to efficiently identify feasible solutions.

NSGA-III [5], an evolutionary algorithm that can tackle multi-objective optimization problems, was adopted to solve DJSS problems with customer time requirements since it has good performance even for problems with more than two objectives [11,12,15].

Each chromosome in the population is represented as a permutation of length $N + \sum_{u=1}^{F} M_u - 1$ genes, encoding a feasible sequence of jobs in machines in different factories (π). Within the chromosome, genes with values greater than N act as delimiters that partition the permutation into subsequences ($\pi_{u,i}$). Each subsequence specifies the set of jobs, together with their processing order, assigned to machine $i \in \mathbb{M}_u$ in the factory $u \in \mathbb{F}$.

To select chromosomes for reproduction, a binary tournament selection operator was used, allowing offspring to be generated from parent solutions through the application of genetic operators. To ensure solution feasibility in the search process, order-based crossover and inversion mutation were adopted. In order-based crossover, genes from the two parent chromosomes are combined by first

selecting two random positions and swapping the genes located between them. The remaining empty positions are then filled with the genes from the other parent while avoiding duplicates. In the inversion mutation, two positions are randomly chosen, and the sequence of genes between them is reversed.

As result of the optimization process, an approximation to the Pareto-optimal set is obtained. Each solution in this set is a different trade-off in terms of the criteria. Various tools can be used to present and visualize a solution, such as Gantt charts to display job scheduling, graph-based representations of the Pareto-optimal frontier, or parallel coordinate plots to highlight trade-offs among objectives. These visualization techniques are intended to support a clearer analysis of solutions to assist in the decision-making process.

3 Experimental Results

3.1 General Setup

In order to test the approach to solve distributed job shop scheduling problems with customer time requirements, three instances (P_1, P_2, and P_3) are borrowed from a previous work [16]. The P_1 instance has 6 jobs ($N = 6$) to be assigned to machines distributed by 2 factories ($F = 2$). Each factory has 2 machines ($M_1 = 2$ and $M_2 = 2$). For this instance, the delivery deadlines for the jobs, reflecting customer time requirements, are $d_1 = \big(200, 230, 200, 260, 285, 235\big)^T$.

Instances P_2 and P_3 both have 11 jobs ($N = 11$). The P_2 instance has 2 factories ($F = 2$) but with different numbers of machines, one with 2 machines and the other with 4 machines ($M_1 = 2$ and $M_2 = 4$). Finally, instance P_3 has 3 factories ($F = 3$), two of them with 2 machines, and the other with 4 machines ($M_1 = 2$, $M_2 = 2$ and $M_3 = 4$). Additionally, the delivery deadlines for the jobs for instances P_2 and P_3 are $d_2 = \big(200, 230, 200, 260, 285, 235, 205, 290, 246, 278, 201\big)^T$ and $d_3 = \big(199, 200, 230, 296, 200, 201, 260, 285, 235, 188, 268\big)^T$, respectively. Processing times p, setup times s, and transportation times t are given in a previous study [16].

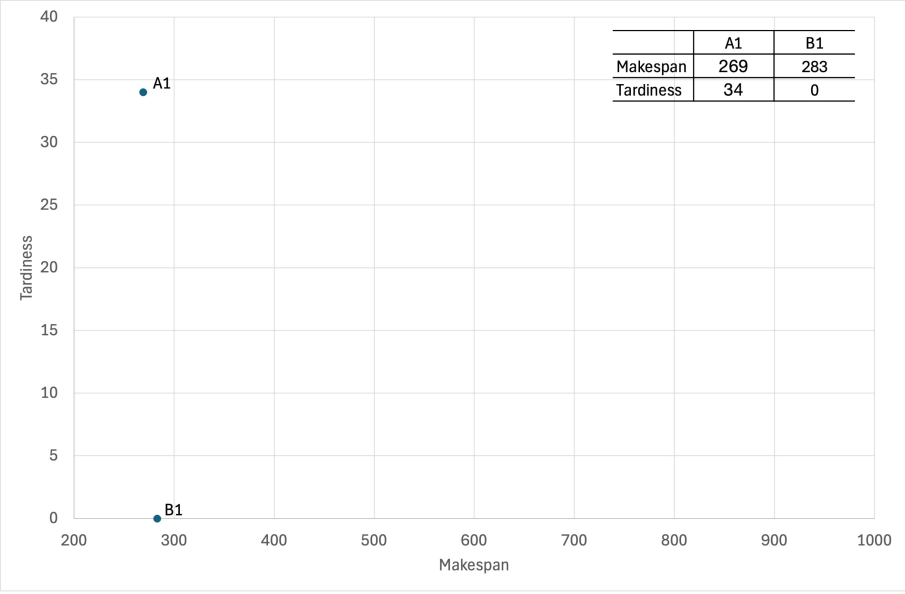
In this study, NSGA-III was used to assess the proposed multi-objective DJSS with the customer time requirements approach. All instances were solved considering the DJSS problems $\mathbf{F_1}(\pi)$, $\mathbf{F_2}(\pi)$, and $\mathbf{F_3}(\pi)$ expressed by the equations (4), (5), and (6), respectively. The implementation was carried out in Python using the PyMOO framework [2]. The population size was set to 100 chromosomes for bi-objective problems $\mathbf{F_1}(\pi)$ and $\mathbf{F_2}(\pi)$, and 200 chromosomes for problem $\mathbf{F_3}(\pi)$ with 3 objectives. The stopping criteria combined a maximum generation limit with a stagnation threshold: the algorithm terminated after reaching 10,000 generations or after 100 consecutive generations without improvement. In addition, a limit of 500,000 objective function evaluations was imposed.

Due to the stochastic nature of evolutionary algorithms, 20 independent runs were conducted for each instance and problem. In the following, with respect to

each instance, the results for the run with greater hypervolume (among the 20 independent runs) are reported. Note that the non-dominated solutions obtained in the run with the highest hypervolume value correspond to the best approximation to the Pareto optimal set. The reference point for hypervolume computation was $(2000, 2000)^T$ for bi-objective problems ($\mathbf{F_1}(\pi)$ and $\mathbf{F_2}(\pi)$). For problem $\mathbf{F_3}(\pi)$ with 3 objectives, the reference point $(2000, 2000, 2000)^T$ was considered.

3.2 P_1 Instance Analysis

In Fig. 1, for instance P_1, the approximation to the Pareto optimal front obtained when solving problem $\mathbf{F_1}(\pi)$, the minimization of makespan and tardiness, is represented. This Pareto front intends to facilitate the decision-maker process, i.e., to select the most convenient solution according to his/her preferences.



	A1	B1
Makespan	269	283
Tardiness	34	0

Fig. 1. Representation of Pareto optimal front for the minimization of makespan and tardiness ($\mathbf{F_1}(\pi)$ problem) for the instance P_1.

For this instance, in all 20 independent runs of NSGA-III, the same two non-dominated solutions (solutions A1 and B1) were obtained. This figure also includes the pay-off table with the values of the objectives for each solution. Solution A1 is better than solution B1 in terms of makespan, but worst in terms of tardiness. Actually, solution A1 has a tardiness value superior to zero, which means that customer delivery dates for the jobs are not fulfilled completely. On

the other hand, solution B1 satisfies all delivery dates with a slightly higher value of makespan. So, it is worth taking a deeper analysis of these solutions using Gantt diagrams to represent the corresponding scheduling of jobs.

For illustration purposes, Fig. 2 and Fig. 3 are Gantt diagrams, representing the scheduling of jobs corresponding to solutions A1 and B1, for instance P_1 with 2 factories, both with 2 machines ($m_{1,1}$, $m_{1,2}$, $m_{2,1}$ and $m_{2,2}$) and 6 jobs (n_1, n_2, n_3, n_4, n_5 and n_6). In these figures, the transportation times, setup times, and processing times of jobs are represented by gray, blue, and orange boxes. The customer delivery dates for jobs (d_1, d_2, d_3, d_4, d_5 and d_6) are also indicated in the diagram.

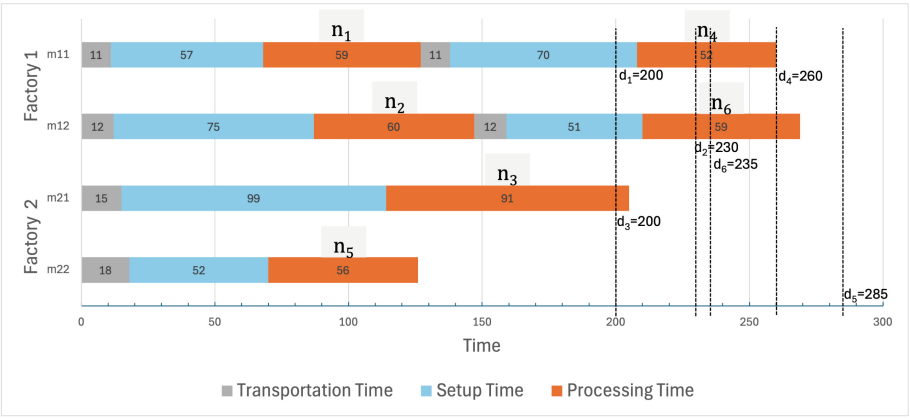


Fig. 2. Gantt diagram representing solution A1 for problem $\mathbf{F_1}(\pi)$ for instance P_1.

The Gantt diagram (Fig. 2) represents the solution A1 and corresponds to the sequence $\pi =< \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2} >$ where $\pi_{1,1} = [1, 4]$, $\pi_{1,2} = [2, 6]$, $\pi_{2,1} = [3]$ and $\pi_{2,2} = [5]$. The total completion times are $c_{1,1}(\pi) = 260$, $c_{1,2}(\pi) = 269$, $c_{2,1}(\pi) = 205$ and $c_{2,2}(\pi) = 126$, and therefore $c_{\max}(\pi) = 269$. Since $c^1_{1,1}(\pi) = 127$, $c^4_{1,1}(\pi) = 260$, $c^2_{1,2}(\pi) = 147$, $c^6_{1,2}(\pi) = 269$, $c^3_{2,1}(\pi) = 205$ and $c^5_{2,2}(\pi) = 126$, then $t_{1,1}(\pi) = 0$, $t_{1,2}(\pi) = 34$, $t_{2,1}(\pi) = 5$, and $t_{2,2}(\pi) = 0$ and therefore $t_{\max}(\pi) = 34$. Note that, despite this solution having the same makespan, it is different from solution A previously published in [16] since the order of jobs in machine 1 of factory 1 (job n_1 before job n_4) is the opposite due to the existence of customer delivery dates.

Solution B1 is represented in the Gantt diagram (Fig. 3). This solution corresponds to the sequence $\pi =< \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2} >$ where $\pi_{1,1} = [1, 4]$, $\pi_{1,2} = [3]$, $\pi_{2,1} = [2]$ and $\pi_{2,2} = [6, 5]$. The total completion times are $c_{1,1}(\pi) = 260$, $c_{1,2}(\pi) = 164$, $c_{2,1}(\pi) = 168$ and $c_{2,2}(\pi) = 283$, and therefore $c_{\max}(\pi) = 283$. Since $c^1_{1,1}(\pi) = 127$, $c^4_{1,1}(\pi) = 260$, $c^3_{1,2}(\pi) = 164$, $c^2_{2,1}(\pi) = 168$, $c^6_{2,2}(\pi) = 158$ and $c^5_{2,2}(\pi) = 283$, then $t_{1,1}(\pi) = 0$, $t_{1,2}(\pi) = 0$, $t_{2,1}(\pi) = 0$, and $t_{2,2}(\pi) = 0$ and

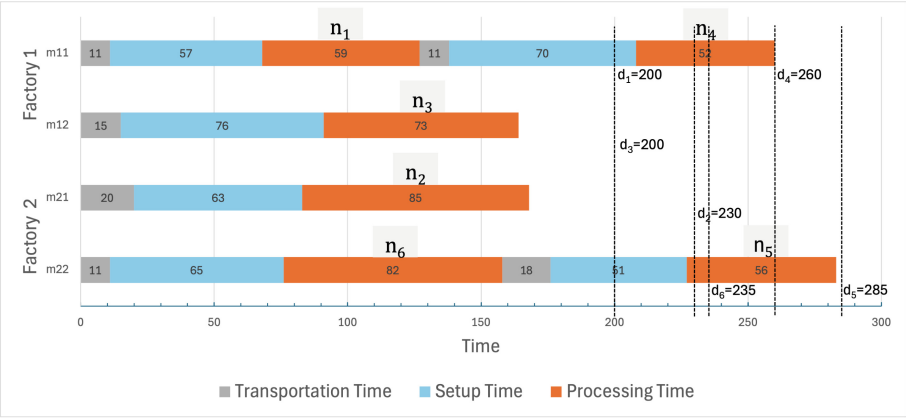


Fig. 3. Gantt diagram representing solution B1 for problem $\mathbf{F_1}(\pi)$ for instance P_1.

therefore $t_{\max}(\pi) = 0$. This solution satisfies all delivery dates for the customer. Thus, it is a good solution from the point of view of the customer's requirements.

In Fig. 4, for instance P_1, the approximation to the Pareto optimal front obtained when solving problem $\mathbf{F_2}(\pi)$, the minimization of makespan and earliness, is represented. Four non-dominated solutions (A2, B2, C2 and D2) were found. Solutions A2 and D2 are extreme solutions located at the upper-left and lower-right extremes of the Pareto front, respectively.

Although solution A2 has the same makespan value as solution A1 of problem $\mathbf{F_1}(\pi)$, it corresponds to a different scheduling. Figure 5 shows the scheduling of jobs corresponding to the sequence $\pi =< \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2} >$ where $\pi_{1,1} = [1, 4]$, $\pi_{1,2} = [2, 6]$, $\pi_{2,1} = [5]$ and $\pi_{2,2} = [3]$. This solution is very similar to solution A1, but with jobs n_3 and n_5 assigned to different machines in factory 2 (Fig. 5).

The total completion times for solution A2 are $c_{1,1}(\pi) = 260$, $c_{1,2}(\pi) = 269$, $c_{2,1}(\pi) = 185$, and $c_{2,2}(\pi) = 174$, and therefore $c_{\max}(\pi) = 269$. Since $c^1_{1,1}(\pi) = 127$, $c^4_{1,1}(\pi) = 260$, $c^2_{1,2}(\pi) = 147$, $c^6_{1,2}(\pi) = 269$, $c^5_{2,1}(\pi) = 185$ and $c^3_{2,2}(\pi) = 174$, then $l_{1,1}(\pi) = 73$, $l_{1,2}(\pi) = 83$, $l_{2,1}(\pi) = 100$, and $l_{2,2}(\pi) = 26$ and therefore $l_{\max}(\pi) = 100$. Note that this solution A2 has the same tardiness value ($t_{\max}(\pi) = 34$) as that of solution A1. However, the earliness of solution A1 is higher ($l_{\max}(\pi) = 159$). Therefore, solution A1 is dominated by solution A2 when the 3 objectives are considered simultaneously.

Solutions B2 and C2 are located at the elbow of the Pareto front and can be good trade-offs between both objectives. Its earliness values do not differ significantly. At the other extreme of the Pareto front, solution D2 has zero earliness, which means that all jobs terminate exactly at the customer delivery date, but with a high makespan. The sequence $\pi =< \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2} >$ for solution D2 is $\pi_{1,1} = [\,]$, $\pi_{1,2} = [\,]$, $\pi_{2,1} = [3, 5, 2, 1, 6, 4]$ and $\pi_{2,2} = [\,]$ with a

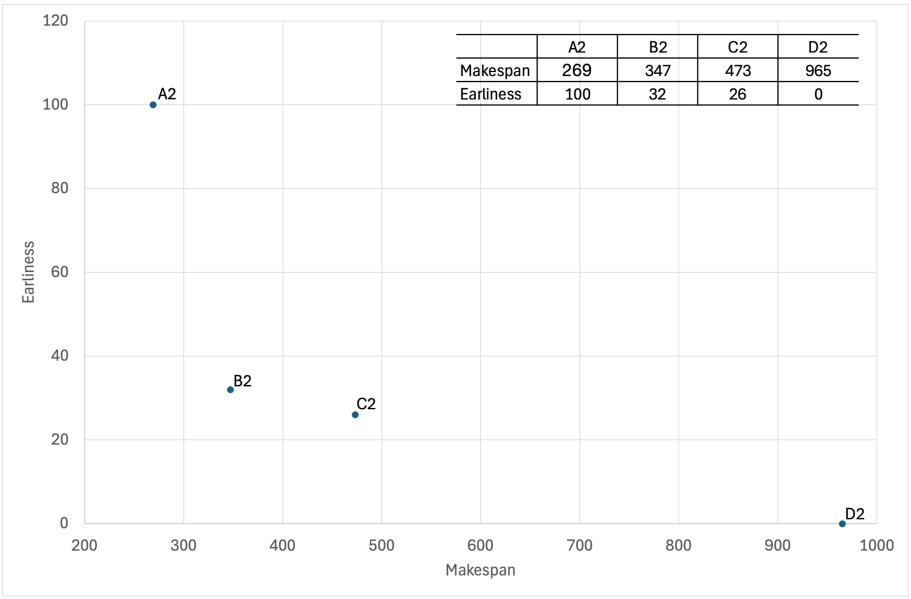


Fig. 4. Representation of Pareto optimal front for the minimization of makespan and earliness ($\mathbf{F_2}(\pi)$ problem) for the instance P_1.

makespan of $c_{\max}(\pi) = 965$. In this solution D2, all jobs are assigned to machine 1 of factory 2. All other machines are not occupied.

The results of the problems $\mathbf{F_1}(\pi)$ and $\mathbf{F_2}(\pi)$ highlight the importance of simultaneously considering makespan, tardiness, and earliness. So, in Fig. 6, the approximation to the Pareto optimal front obtained for the problem $\mathbf{F_3}(\pi)$ is shown. In this figure, makespan, tardiness, and earliness are represented on the X-axis, the Y-axis, and by the size of the points. It can be observed that a larger number of non-dominated solutions was found compared to the bi-objective formulations ($\mathbf{F_1}(\pi)$ and $\mathbf{F_2}(\pi)$). As stated above, the solution A3 is equal to A2, but different from A1. On the other hand, solutions B3, F3, G3 and H3 are equal to solutions B1, B2, C2 and D2, respectively.

From Fig. 6, it seems that, in general, makespan and earliness are negatively associated, i.e., when makespan increases, earliness tends to decrease. This behavior does not occur with tardiness. These results confirm the conflicting nature of these three objectives.

3.3 P_2 and P_3 Instances Analysis

The problems $\mathbf{F_1}(\pi)$, $\mathbf{F_2}(\pi)$, and $\mathbf{F_3}(\pi)$ were solved for instances P_2 and P_3. Since all nom-dominated solutions obtained for $\mathbf{F_1}(\pi)$ and $\mathbf{F_2}(\pi)$ were also obtained when solving $\mathbf{F_3}(\pi)$, in this paper, only the results for the latter problem are presented.

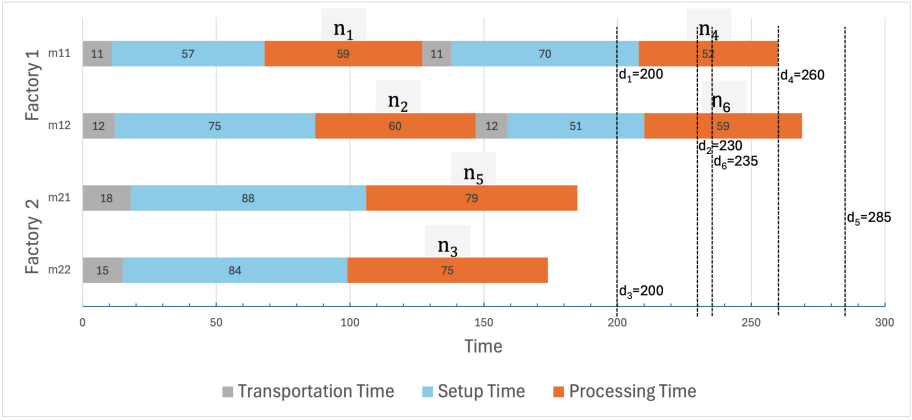


Fig. 5. Gantt diagram representing solution A2 for problem $\mathbf{F_2}(\pi)$ for instance P_1.

In Fig. 7, the approximation to the Pareto optimal front obtained for the minimization of the makespan, tardiness and earliness for the instance P_2 is shown. For this instance, 11 non-dominated solutions were found (solutions A to K). Solutions A, B and K are the solutions with extreme values of the objectives. It should be noted that there is no solution with a tardiness equal to zero, i.e., the customer delivery dates cannot be all satisfied simultaneously. The solution with the smallest tardiness has a delay of 24. No solution was found with earliness equal to zero. However, solution K has an earliness value very small (equal to one). These solutions A, B and K correspond to the scheduling plans defined by the following sequences $\pi =< \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2}, \pi_{2,3}, \pi_{2,4} >$:

- Solution A: $\pi_{1,1} = [1, 10]$, $\pi_{1,2} = [7]$, $\pi_{2,1} = [4, 8]$, $\pi_{2,2} = [3, 2]$, $\pi_{2,3} = [11, 9]$, and $\pi_{2,4} = [5, 6]$;
- Solution B: $\pi_{1,1} = [2, 10, 5]$, $\pi_{1,2} = [11, 6, 8]$, $\pi_{2,1} = [3, 4, 1]$, $\pi_{2,2} = [\,]$, $\pi_{2,3} = [7, 9]$, and $\pi_{2,4} = [\,]$;
- Solution K: $\pi_{1,1} = [\,]$, $\pi_{1,2} = [\,]$, $\pi_{2,1} = [\,]$, $\pi_{2,2} = [\,]$, $\pi_{2,3} = [3, 1, 9, 2, 7, 8, 11, 10, 5, 6, 4]$, and $\pi_{2,4} = [\,]$.

Solution A has the same makespan as solution A previously reported in [16] for the same instance P_2. Despite sharing some similarities, they have different jobs assigned to machines in factory 1. Other alternative solutions representing different trade-offs between objectives were also found.

The number of non-dominated solutions is larger due to the combinatorial nature of the problem and the conflicting objectives. We are reporting the approximation to the Pareto optimal front obtained from the run with the best hypervolume. However, it should be noted that taking into account all approximations to the Pareto optimal front obtained on the 20 runs, 36 non-dominated solutions were identified for the instance P_2. This suggests the existence of a higher number of solutions in the Pareto optimal front.

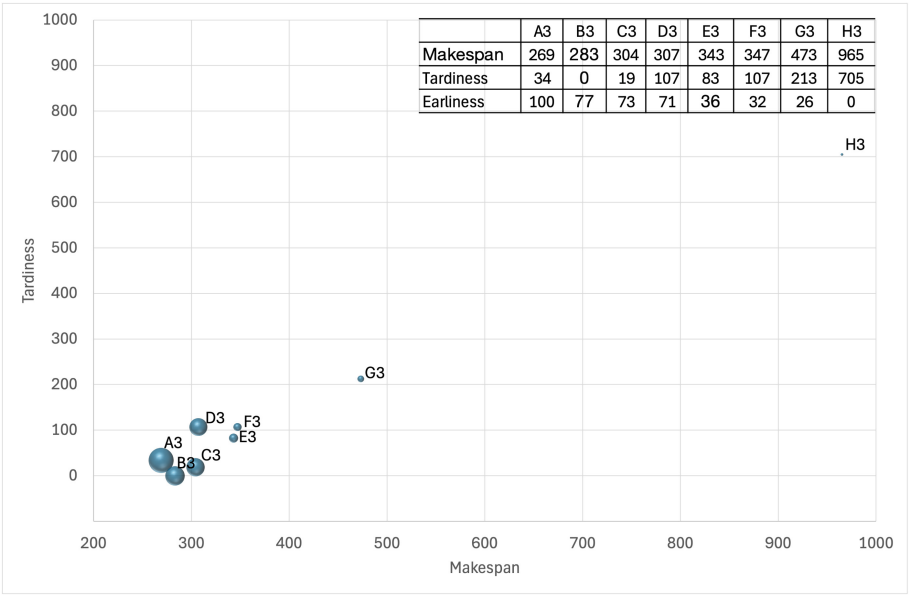


Fig. 6. Representation of Pareto optimal front for minimization of makespan, tardiness and earliness ($\mathbf{F_3}(\pi)$ problem) for the instance P_1.

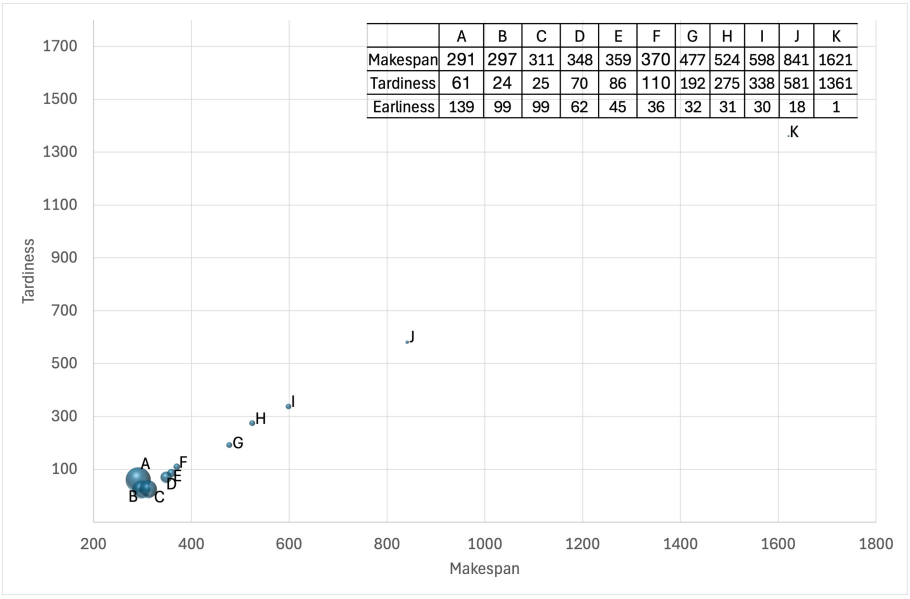


Fig. 7. Representation of Pareto optimal front for minimization of makespan, tardiness and earliness ($\mathbf{F_3}(\pi)$ problem) for the instance P_2.

Figure 8 depicts the approximation to the Pareto optimal front obtained for the minimization of the makespan, tardiness and earliness for the instance P_3. For this instance, the approximation to the Pareto optimal front is constituted by 16 non-dominated solutions (solutions A to P). Solutions A, D and P are the solutions with extreme values of the objectives. For this instance, there exists one solution with tardiness equal to zero (solution D) and another one with earliness equal to zero (solution P). The scheduling plans for these solutions are defined by the sequences $\pi =< \pi_{1,1}, \pi_{1,2}, \pi_{2,1}, \pi_{2,2}, \pi_{3,1}, \pi_{3,2}, \pi_{3,3}, \pi_{3,4} >$:

- Solution A: $\pi_{1,1} = [7,3]$, $\pi_{1,2} = [8]$, $\pi_{2,1} = [11]$, $\pi_{2,2} = [2]$, $\pi_{3,1} = [1]$, $\pi_{3,2} = [5,4]$, $\pi_{3,3} = [6]$, and $\pi_{3,4} = [10]$;
- Solution D: $\pi_{1,1} = [6]$, $\pi_{1,2} = [1,8]$, $\pi_{2,1} = [9,11]$, $\pi_{2,2} = [2]$, $\pi_{3,1} = [3]$, $\pi_{3,2} = [5]$, $\pi_{3,3} = [10,4]$, and $\pi_{3,4} = [7]$;
- Solution P: $\pi_{1,1} = [\,]$, $\pi_{1,2} = [10,4,3,11]$, $\pi_{2,1} = [\,]$, $\pi_{2,2} = [\,]$, $\pi_{3,1} = [2,5,9]$, $\pi_{3,2} = [\,]$, $\pi_{3,3} = [\,]$, and $\pi_{3,4} = [6,1,7,8]$.

In Fig. 8, the approximation to the Pareto optimal front obtained from the run with the best hypervolume is depicted. However, taking into account all approximations to the Pareto optimal front obtained in the 20 runs, 34 non-dominated solutions were identified for this instance P_3. Therefore, as for instance P_2, it seems that, due to the combinatorial nature of the search space, the Pareto optimal set has a higher cardinality.

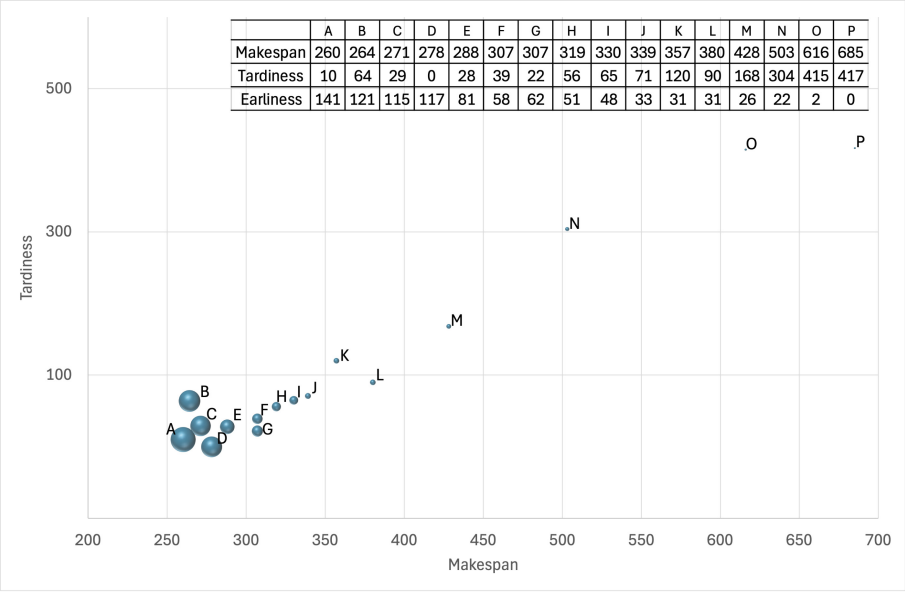


Fig. 8. Representation of Pareto optimal front for minimization of makespan, tardiness and earliness ($\mathbf{F_3}(\pi)$ problem) for the instance P_3.

From Fig. 7 and Fig. 8, it can observed again the existence of a tendency similar to that identified for the instance P_1, i.e., in general, makespan and earliness are negatively associated. However, this tendency does not exist with tardiness.

These results suggest that this approach to solving DJSS problems with customer delivery date requirements provides the decision-maker with relevant information about scheduling plans from both production and customer satisfaction perspectives.

4 Conclusions and Future Work

This paper proposes a multi-objective optimization approach for solving distributed job shop scheduling (DJSS) problems in an environment with unrelated parallel machines and sequence-dependent setup times. The main contribution of this study lies in the incorporation of customer preferences into the model, achieved through the simultaneous optimization of three fundamental objectives: minimization of makespan ($c_{\max}$), minimization of tardiness ($t_{\max}$) and minimization of earliness ($l_{\max}$). To this effect, the NSGA-III was employed, recognized for its effectiveness in addressing multi-objective problems. To this end, three test instances (P_1, P_2, and P_3) were considered, with different dimensions and configurations adapted from the previous work incorporating customer delivery dates for jobs.

In instance P_1, it was found that a diverse set of non-dominated solutions could be identified for the two-objective problems F_1 and F_2 , as well as solutions for the three-objective problem F_3. In instances P_2 and P_3, approximations to the Pareto optimal solutions were obtained. However, the results suggest that these Pareto optimal sets include a large number of solutions due to the combinatorial nature of the problems.

The graphical representation of the obtained approximations to the Pareto optimal fronts and the analysis of Gantt charts for specific solutions revealed the trade-offs inherent in the different criteria. It was observed that minimizing the makespan may lead to missed deadlines (tardiness), while strict compliance with delivery dates may require an increase in total production time. On the other hand, minimizing earliness can help reduce costs associated with customer inventory. These results demonstrate the usefulness of the proposed approach in supporting the decision-making process, providing the decision-maker with a set of optimal alternatives that reflect different trade-offs between production efficiency and the level of customer service.

Future work intends to extend the study to larger and more complex systems, as well as to explore the incorporation of additional objectives, such as reducing the number of idle machines or considering environmental sustainability criteria. Finally, comparing the performance of NSGA-III with other multi-objective metaheuristics is also a relevant line of research for future developments.

References

1. Arnaout, J.P., Musa, R., Rabadi, G.: A two-stage ant colony optimization algorithm to minimize the makespan on unrelated parallel machines—Part II: enhancements and experimentations. J. Intell. Manuf. **25**, 43–53 (2014)
2. Blank, J., Deb, K.: PYMOO: multi-objective optimization in python. IEEE Access **8**, 89497–89509 (2020)
3. Chan, F.T., Chung, S.H., Chan, L., Finke, G., Tiwari, M.: Solving distributed FMS scheduling problems subject to maintenance: genetic algorithms approach. Robot. Comput.-Integr. Manuf. **22**(5–6), 493–504 (2006)
4. De Giovanni, L., Pezzella, F.: An improved genetic algorithm for the distributed and flexible job-shop scheduling problem. Eur. J. Oper. Res. **200**(2), 395–408 (2010)
5. Deb, K., Jain, H.: An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints. IEEE Trans. Evol. Comput. **18**(4), 577–601 (2014)
6. Khan, B., Hanoun, S., Johnstone, M., Lim, C.P., Creighton, D., Nahavandi, S.: Multi-objective job shop scheduling using I-NSGA-III. In: 2018 Annual IEEE International Systems Conference (SysCon), pp. 1–5. IEEE (2018)
7. Lohmer, J., Lasch, R.: Production planning and scheduling in multi-factory production networks: a systematic literature review. Int. J. Prod. Res. **59**(7), 2028–2054 (2021)
8. Marzouki, B., Driss, O.B., Ghédira, K.: Solving distributed and flexible job shop scheduling problem using a chemical reaction optimization metaheuristic. Procedia Comput. Sci. **126**, 1424–1433 (2018)
9. Santos, F., Costa, L.: Multivariate analysis to assist decision-making in many-objective engineering optimization problems. In: Gervasi, O., et al. (eds.) ICCSA 2020. LNCS, vol. 12251, pp. 274–288. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58808-3_21
10. dos Santos, F., Costa, L., Varela, L.: Multi-objective optimization of the job shop scheduling problem on unrelated parallel machines with sequence-dependent setup times. In: Gervasi, O., et al. (eds.) ICCSA 2023. LNCS, vol. 14105, pp. 495–507. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37108-0_32
11. dos Santos, F., Costa, L., Varela, L.: Multi-objective scheduling optimization in job shop with unrelated parallel machines using NSGA-III. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., C. Rocha, A.M.A., Faginas Lago, M.N. (eds.) ICCSA 2024. LNCS, vol. 14816, pp. 370–382. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-65223-3_25
12. dos Santos, F., Costa, L., Varela, L.: Analysis of the effect of global and local objectives in multi-objective distributed job shop scheduling problems. In: Gervasi, O., et al. (eds.) ICCSA 2025. LNCS, vol. 15888, pp. 180–193. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-97596-7_12
13. dos Santos, F., Costa, L.A., Varela, L.: A systematic literature review about multi-objective optimization for distributed manufacturing scheduling in the industry 4.0. In: Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C. (eds.) ICCSA 2022, Part II. LNCS, vol. 13378, pp. 157–173. Springer (2022). https://doi.org/10.1007/978-3-031-10562-3_12
14. dos Santos, F., Costa, L.A., Varela, L.: Multiobjective optimization in distributed industry and environmental sustainability: a systematic literature review. Revista Angolana de Ciências **5**(2), e050210 (2024)

15. dos Santos, F., Costa, L.A., Varela, L.: Performance comparison of NSGA-II and NSGA-III on bi-objective job shop scheduling problems. In: Pereira, A.I., Mendes, A., Fernandes, F.P., Pacheco, M.F., Coelho, J.P., Lima, J. (eds.) OL2A 2023. LNCS, vol. 1981, pp. 531–543. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-53025-8_36
16. Santos, F.d., Costa, L., Varela, L.: A multi-objective approach for solving distributed job shop scheduling problems. In: Pereira, A.I., et al. (eds.) OL2A 2024. . LNCS, vol. 2280, pp. 326–339. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-77426-3_22
17. Yenisey, M.M., Yagmahan, B.: Multi-objective permutation flow shop scheduling problem: literature review, classification and current trends. Omega **45**, 119–135 (2014)
18. Zhang, J., Ding, G., Zou, Y., Qin, S., Fu, J.: Review of job shop scheduling research and its new perspectives under industry 4.0. J. Intell. Manuf. **30**, 1809–1830 (2019)

Optimizing Directional Change Trading Strategies with Adaptive Thresholds for the Brazilian Stock Market

Lucas E. Mieri[1(✉)] and Sanderson L. Gonzaga de Oliveira[2]

[1] Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, Brazil
lucasmieri@ita.br
[2] Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil
sanderson.oliveira@unifesp.br

Abstract. Financial time series analysis has traditionally relied on physical time, where analysts sample prices at fixed intervals. The Directional Change (DC) paradigm replaces this with an event-based representation driven by price magnitude rather than elapsed time. We extend the Multi-Strategy/Threshold Genetic Algorithm Model, which combines eight DC-based trading strategies across multiple thresholds into an ensemble of 70 sub-strategies optimized by a Genetic Algorithm (GA), by introducing volatility-adaptive thresholds that scale to each asset, walk-forward validation with per-fold recalibration, and the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) as an alternative optimizer. This study evaluates the framework on 30 equities from the Brazilian stock exchange across 270 out-of-sample folds spanning from 2017 to 2025. The CMA-ES ensemble achieves a higher Sharpe Ratio than Buy-and-Hold (0.87 vs. 0.64) and nearly doubles its profit factor, while reducing volatility in every fold tested (mean ratio 0.68). CMA-ES and GA achieve comparable out-of-sample performance, but CMA-ES generalizes better, with half the train-to-test Sharpe degradation.

Keywords: Directional Changes · Genetic Algorithm · CMA-ES · Trading Strategies · Brazilian Stock Market · Walk-Forward Validation

1 Introduction

Analysts have largely confined financial time series analysis to physical time, sampling prices at fixed intervals such as daily closes or minute-by-minute snapshots [1,2]. This convention is practical but lossy: price movements that fall between two consecutive sampling points go unrecorded, and can miss significant events [3]. The Directional Change (DC) paradigm offers an event-based alternative. Introduced by Guillaume et al. [1] and later formalized through the scaling laws of Glattfelder et al. [4], DC decomposes a price series into alternating upward and downward movements defined by a threshold θ. A reversal of at least θ from the last extreme confirms a DC event; the continuation beyond that

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 199–214, 2027.
https://doi.org/10.1007/978-3-032-30521-3_13

confirmation point until the next reversal is called the Overshoot (OS). Because events are triggered by price magnitude rather than elapsed time, the resulting representation captures only movements that surpass the threshold, filtering out small fluctuations that may not carry directional information [3].

Salman et al. [5] proposed the Multi-Strategy/Threshold Genetic Algorithm Model (MSTGAM), which combines eight DC-based trading strategies across ten thresholds into an ensemble of 70 sub-strategies with weights optimized by a Genetic Algorithm (GA). Each strategy exploits a distinct statistical regularity of large price movements, so that the ensemble extracts predictive information from the events that the DC paradigm isolates as significant. The model reported superior risk-adjusted returns over both individual DC strategies and Technical Analysis (TA) benchmarks on the New York Stock Exchange (NYSE). However, the authors designed the framework for NYSE data, and several choices limit its applicability to other markets. The thresholds (0.098% to 2.55%) are hard-coded for developed-market volatility. Evaluation relies on a single 80/20 split, with little information about robustness across time. The model fixes the risk-free rate at 2.5%, far below the Brazilian Interbank Deposit Certificate (CDI), the standard risk-free benchmark in Brazil, which has ranged from 2% to 14% over the past decade. Strategy parameters are calibrated once on the full training set, introducing the risk of look-ahead bias; thus, any strategy with genuine predictive capacity should exhibit low train-to-test Sharpe degradation, defined as the difference between Sharpe ratios on the training and test sets. The GA itself requires 50 independent runs per configuration [6] to compensate for its stochastic variability, and CMA-ES [7], widely regarded as state-of-the-art for continuous black-box optimization [8,9], has not been applied to DC ensemble weights.

Three open problems motivate the present work. First, all existing DC trading studies employ fixed thresholds derived manually or from a predefined grid; no prior work adapts thresholds to asset-specific volatility, limiting cross-market applicability. Second, walk-forward evaluation with per-fold recalibration remains uncommon in DC trading studies, where single train-test splits are the prevailing practice. Third, analysts have not applied CMA-ES [7,8] to DC strategy weights, with only isolated uses in computational finance [19].

Notably, analysts have studied all existing DC trading in markets from advanced economies (e.g., FX pairs, NYSE, London Stock Exchange), as well as the Indian stock market. The Brazilian market, although being the largest in Latin America, has not been explored with MSTGAM or any other DC-based trading framework.

This work extends the MSTGAM framework with the following contributions:

1. **Volatility-adaptive threshold calibration**: the model scales DC thresholds to each asset's realized daily volatility; consequently, the same framework applies across assets without manual tuning.

2. **Walk-forward validation with per-fold calibration**: each strategy parameter is recalibrated at each fold using only training data, preventing any future information from entering the optimization.
3. **CDI-adjusted evaluation**: each performance metric employs the daily CDI as the risk-free rate, providing a fair comparison in the Brazilian context.
4. **CMA-ES as an ensemble optimizer**, replacing the multi-run GA with a single optimization per fold.
5. **First application of the DC paradigm to the Brazilian stock exchange (B3)**: this study evaluates the framework on 30 equities spanning multiple sectors over the period from 2015 to 2025.

Based on these extensions, we investigate four hypotheses:

H1. The MSTGAM ensemble with volatility-adaptive thresholds achieves a higher Sharpe Ratio [10] than Buy-and-Hold on B3 equities under out-of-sample walk-forward evaluation.
H2. The DC-based ensemble systematically reduces portfolio volatility and maximum drawdown (MDD) relative to the Buy-and-Hold strategy, operating as a risk filter (a strategy that primarily reduces volatility and drawdown) rather than an alpha generator (a strategy that primarily increases returns above the benchmark).
H3. The advantage of the ensemble over Buy-and-Hold increases with the size of the training window, as exposure to multiple market regimes improves strategy calibration.
H4. CMA-ES exhibits lower train-to-test Sharpe degradation than the GA, indicating better generalization to unseen data.

Section 2 reviews related work on DC-based trading and metaheuristic optimization. Section 3 presents the DC detection algorithm and the eight trading strategies that form the base of the ensemble. Section 4 describes the proposed extensions: volatility-adaptive thresholds, the ensemble optimization with CMA-ES and GA, and the fitness function. Section 5 details the experimental methodology. Section 6 presents and discusses the results. Section 7 concludes the paper.

2 Related Work

The DC paradigm originates from the work of Guillaume et al. [1], who proposed an event-driven approach to summarize price movements in foreign exchange (FX) markets. Glattfelder et al. [4] subsequently identified 12 empirical scaling laws from high-frequency FX data using this framework. Aloud et al. [3] formalized the detection algorithm that decomposes a price series into alternating DC and OS events, given a threshold θ, and Tsang et al. [2] extended the application of DC indicators to equity markets. Tao [11] developed additional DC-based indicators to extract information from financial data. Salman et al. [12] introduced the optimization of DC-based trading strategies using GA, combining seven rule-based strategies on a single threshold into the Multi-Strategy GA Model

(MSGAM). In a subsequent work, Salman et al. [13] extended the approach by applying GA optimization over multiple thresholds for each strategy individually, known as the Multi-Threshold GA Model (MTGAM). The full integration of all eight strategies and thresholds into a single ensemble of 70 sub-strategies, the MSTGAM, was presented in Salman et al. [5], reporting improved Sharpe Ratios over individual strategies and TA benchmarks on 200 NYSE stocks. In short, MSGAM combines multiple strategies on a single threshold, MTGAM optimizes individual strategies across multiple thresholds, and MSTGAM unifies both dimensions into a single ensemble of strategy-threshold pairs.

Beyond GA-based approaches, analysts have applied other metaheuristics to DC trading. Long et al. [14] compared Genetic Programming (GP) under both physical time and DC frameworks, finding that DC-based GP strategies outperform their physical-time counterparts on FX data. Long et al. [15] proposed a multi-objective GP approach that integrates DC signals with a modified Sharpe Ratio as fitness function, using the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to address the single-objective limitation of earlier models. Rayment and Kampouridis [16] applied Deep Reinforcement Learning (DRL) using Proximal Policy Optimization to high-frequency FX trading within a DC sampling framework, and later extended the approach with position-sensitive mechanisms to enhance state representation [17]. Outside the DC paradigm, Chinraj and Anthonysamy [18] combined a GA with the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) in a rule-based moving average strategy to optimize the Indian stock market, using walk-forward validation.

3 Background: DC Paradigm and Trading Strategies

The ensemble relies on two building blocks: a DC detection algorithm and eight trading strategies. We first present the DC detection algorithm in Sect. 3.1 and then summarize the eight trading strategies that serve as the base components in Sect. 3.2.

3.1 Directional Change Detection

As outlined in Sect. 2, the DC paradigm replaces fixed-interval sampling with an event-driven representation in which the price series is dissected solely by the magnitude of price reversals. Given an adjusted close price series $\{p_t\}_{t=1}^{N}$ and a threshold $\theta > 0$, the detection algorithm (Algorithm 1) maintains a running extreme price and the current trend mode. A downturn (DT) is confirmed when the price falls by at least θ from the running high, and an upturn (UT) when it rises by at least θ from the running low. The interval between the extreme point and its confirmation defines the DC event; the continuation beyond the DC Confirmation (DCC) point until the next reversal defines the OS event. Two empirical scaling laws emerge from this decomposition [4]: the average OS price change approximates the threshold ($\langle \Delta p_{OS} \rangle \approx \theta$), reflecting the symmetry of price movements around threshold-defined reversals; and the average OS

duration approximates twice the DC duration ($\langle T_{OS} \rangle \approx 2\langle T_{DC} \rangle$), reflecting that markets typically take longer to exhaust an established trend than to confirm its reversal.

Algorithm 1: Directional Change detection (adapted from Aloud et al. [3]).

Input: prices $\{p_t\}_{t=1}^{N}$, threshold $\theta > 0$, initial mode $\in \{\text{UT}, \text{DT}\}$
Output: list of DC events, per-bar indicators

$p_{ext} \leftarrow p_1$; $mode \leftarrow$ initial mode;
for $t = 2, \ldots, N$ **do**
  if $mode = UT$ **then**
    if $p_t > p_{ext}$ **then**
      $p_{ext} \leftarrow p_t$;
    else if $(p_{ext} - p_t)/p_{ext} \geq \theta$ **then**
      record DT event;
      $mode \leftarrow$ DT;
      $p_{ext} \leftarrow p_t$;
    end if
  else if $mode = DT$ **then**
    if $p_t < p_{ext}$ **then**
      $p_{ext} \leftarrow p_t$;
    else if $(p_t - p_{ext})/p_{ext} \geq \theta$ **then**
      record UT event;
      $mode \leftarrow$ UT;
      $p_{ext} \leftarrow p_t$;
    end if
  end if
  Compute per-bar indicators: OSV_{cur}, TMV_{cur}, trend, durations;
end for

The algorithm starts by setting the extreme price p_{ext} to the first observation and choosing an initial trend mode. Then iterates over each subsequent bar. In upturn mode, if the current price p_t exceeds p_{ext}, the algorithm updates the extreme upward, tracking the running high. If, instead, the price drops by at least θ relative to p_{ext}, the algorithm confirms a downturn event, switches the mode to DT, and resets p_{ext} to p_t. Downturn mode mirrors this logic: the algorithm tracks the running low and confirms an upturn when the price rises by at least θ from it, switching back to UT and resetting p_{ext}. At the end of each iteration, the algorithm computes the per-bar indicators OSV_{cur}, TMV_{cur}, trend direction, and event durations that the trading strategies consume downstream.

At each bar t, the algorithm computes two indicators used by the trading strategies. The Overshoot Value (OSV), defined as $OSV_{cur} = (p_t - p^*_{DCC})/(\theta \cdot p^*_{DCC})$, measures the current price displacement p_t from the theoretical confirmation point p^*_{DCC}. The Total Move Value (TMV), defined as $TMV_{cur} =$

$(p_t - p_{ext})/(\theta \cdot p_{ext})$, measures the displacement from the last extreme point p_{ext}, encompassing the entire move from that pivot [6]. Additional per-bar state includes the trend direction, the number of bars since the last confirmation, and the duration of the last concluded DC event. These arrays, computed in a single sequential pass over the price series, form the input to all subsequent modules.

3.2 Trading Strategies

Salman et al. [12] assembled eight trading strategies, drawn from DC scaling laws, DC indicators, and TA-inspired saturation patterns, into an ensemble framework. The first group (scaling-law strategies) triggers positions based on empirical DC regularities: one monitors total price displacement relative to the threshold, and the other monitors the duration of the current OS event relative to its preceding DC event. The second group comprises four strategies based on DC indicators and statistical ratios: two compare current values of OSV and TMV against calibrated thresholds, one compares the OS-to-DC duration ratio against a training-set average, and one uses the historical ratio of OS-to-DC event counts as a probabilistic entry filter. The third group detects saturation patterns, defined as sequences of consecutive trends with or without overshoot, analogous to ascending and descending triangles in TA. Table 1 summarizes the entry and exit conditions. Each strategy independently emits a Buy, Hold, or Sell recommendation at every bar, and the model does not permit short selling [6]. Individually, these rules are too simple to sustain consistent returns. The ensemble gains its value from combining its signals: agreement produces conviction, disagreement keeps the portfolio in cash.

Table 1. Entry and exit conditions for the eight DC-based trading strategies (adapted from Salman et al. [12]).

Strategy	Type	Buy Signal (in DT)	Sell Signal
1	Scaling law	$\lvert TMV_{cur}\rvert \geq 2.0$	Same in UT
2	Scaling law	$T_{OS} \geq 2 \cdot T_{DC}$	Same in UT
3	Indicator	$\lvert OSV_{cur}\rvert \geq \lvert OSV_{best}^{DT}\rvert$	$\lvert OSV_{cur}\rvert \geq \lvert OSV_{best}^{UT}\rvert$
4	Indicator	$\lvert TMV_{cur}\rvert \geq \lvert TMV_{best}^{DT}\rvert$	$\lvert TMV_{cur}\rvert \geq \lvert TMV_{best}^{UT}\rvert$
5	Ratio	$T_{OS}/T_{DC} \geq RD$	Same in UT
6	Probabilistic	$r \geq RN$ at DCC	Next DCC in UT
7	Saturation	3 consec. UT with OS	Next DCC in DT
8	Saturation	3 consec. DT with OS	Next DCC in UT

In conceptual terms, Strategy 1 assumes that a trend has exhausted its average magnitude, positioning for a reversal; Strategy 2 applies the same logic to duration rather than price. Strategies 3 and 4 learn from training data the overshoot and total-move levels historically preceded reversals, and trigger when the

current market reaches those levels. Strategy 5 indicates trend exhaustion when the OS phase has lasted disproportionately longer relative to its DC phase. Strategy 6 acts as a probabilistic filter, entering positions more often when overshoots are historically frequent. Strategies 7 and 8 detect repeated failed breakouts in one direction.

4 Adaptive Multi-threshold DC Ensemble for the Brazilian Market

Applying DC-based strategies across assets with different volatility profiles requires thresholds that scale with each asset rather than remaining fixed (Sect. 4.1). Once multiple thresholds produce 70 independent signals, a mechanism is needed to aggregate them into a single trading decision (Sect. 4.2). Finally, the fitness function must reflect the cost and interest-rate environment of the target market (Sect. 4.3).

4.1 Volatility-Adaptive Thresholds

The fixed thresholds used by Salman et al. [5] (0.098% to 2.55%) do not transfer directly to markets with different volatility profiles. For example, BOVA11 has a daily volatility of 1.5% while PETR4 has 2.9% (over the 2015–2025 sample). Applying the same thresholds to both produces excessive event density for the volatile asset and insufficient events for the stable one. Inspired by the logarithmic spacing employed by Glattfelder et al. [4] in their scaling law analysis, we calibrate thresholds proportionally to each asset's realized daily volatility $\hat{\sigma}$, spacing them logarithmically between $f_{\min} \cdot \hat{\sigma}$ and $f_{\max} \cdot \hat{\sigma}$:

$$\theta_i = \exp\left(\ln(f_{\min}\hat{\sigma}) + \frac{i-1}{9}\Big[\ln(f_{\max}\hat{\sigma}) - \ln(f_{\min}\hat{\sigma})\Big]\right), \quad i = 1, \ldots, 10, \tag{1}$$

with $f_{\min} = 0.3$ and $f_{\max} = 5.0$. The logarithmic spacing samples low thresholds more finely, where DC events are most frequent. For BOVA11 ($\hat{\sigma} = 1.49\%$), the resulting range is 0.45% to 7.47%; for PETR4 ($\hat{\sigma} = 2.87\%$), 0.86% to 14.3%. The two saturation-pattern strategies are restricted to the five smallest thresholds, as they generate negligible trades at larger values with daily data [6]. The total number of sub-strategies is therefore $6 \times 10 + 2 \times 5 = 70$.

4.2 Ensemble Aggregation

At each bar t, the 70 sub-strategies independently emit a recommendation $r_j \in \{\text{Buy}, \text{Hold}, \text{Sell}\}$. The ensemble aggregates these into a single decision via stateful weighted voting. Given a weight vector $\mathbf{w} = (w_1, \ldots, w_{70})$ with $w_j \geq 0$ and $\sum_j w_j = 1$, the weighted score for each action is

$$S(a) = \sum_{j=1}^{70} w_j \cdot \mathbf{1}[r_j = a], \quad a \in \{\text{Buy}, \text{Hold}, \text{Sell}\}.$$

Following Salman et al. [5], the approach applies a hold-suppression rule: if at least three sub-strategies emit a non-hold recommendation, the approach sets $S(\text{Hold})$ to zero. The ensemble then updates its position only when Buy or Sell achieves the highest score strictly; if neither dominates, the approach maintains the current position. This inertia mechanism avoids unnecessary round-trips when the ensemble signal is ambiguous.

The weight vector $\mathbf{w}$ is optimized using two metaheuristic approaches, a GA and CMA-ES, described in Sect. 5. Since CMA-ES operates in unconstrained $\mathbb{R}^{70}$, the approach must project the raw output $\mathbf{x}$ onto a valid weight vector. We apply L_1 normalization: $w_j = |x_j| / \sum_k |x_k|$, which guarantees $w_j \geq 0$ and $\sum_j w_j = 1$. For the GA, the approach renormalizes weights so that their sum equals 1 after each crossover and mutation operation.

4.3 Fitness Function

The fitness function evaluates a candidate weight vector by simulating the portfolio it produces over the training period. At each bar t, the portfolio is either fully invested in the asset or fully allocated to the risk-free rate. Transaction costs of 3 basis points per side are deducted at each position change, corresponding to the settlement and trading fees charged by the Brazilian stock exchange for retail equity transactions [20]; Salman et al. [6] report a transaction cost of 0.5 basis points per side for NYSE equities, and the B3 cost is approximately six times higher.

Since the approach executes the trading signal at bar t at the close, the portfolio return during that bar still reflects the previous allocation:

$$r_t^{port} = \begin{cases} r_t^{CDI} - c & \text{if entering long at } t \text{ (still in cash during bar)} \\ r_t^{asset} - c & \text{if exiting long at } t \text{ (still in asset during bar)} \\ r_t^{asset} & \text{if maintaining long position} \\ r_t^{CDI} & \text{if out of market} \end{cases}$$

where $c = 0.0003$ is the one-way transaction cost. The fitness is the annualized calendar-based Sharpe Ratio [10] computed over daily excess returns:

$$SR = \frac{\overline{r^{port} - r^{CDI}}}{\text{std}(r^{port} - r^{CDI})} \cdot \sqrt{252} \tag{2}$$

This approach differs from that of Salman et al. [6], who compute the Sharpe Ratio at the trade level. The calendar-based formulation accounts for the opportunity cost of out-of-market intervals, which the trade-level metric ignores. If the standard deviation of excess returns is close to zero, which occurs when the

weight vector produces no trades, and the portfolio remains entirely in CDI, the Sharpe Ratio is substituted by zero in the corresponding metric to avoid numerical instability (see Sect. 5.3 for details).

5 Methodology

This section describes the walk-forward validation scheme (Sect. 5.1), the data sources and optimizer configurations (Sect. 5.2), and the metrics applied to the out-of-sample results (Sect. 5.3).

5.1 Walk-Forward Validation

To validate the out-of-sample data, we adopt an expanding-window walk-forward protocol. For each fold, the training set comprises all available data from 2015 up to year $Y-1$, and the test set covers year Y. A minimum of two training years is required, yielding nine folds ($Y = 2017, \ldots, 2025$). The first fold trains on 2015–2016 ($\sim$494 bars) and tests on 2017 ($\sim$247 bars); the last fold trains on 2015–2024 ($\sim$2 481 bars) and tests on 2025 ($\sim$248 bars).

At each fold the following steps are executed: (*i*) DC thresholds are computed from the training-set volatility using Eq. 1; (*ii*) the indicator-based and ratio-based strategy parameters are calibrated exclusively on training data; (*iii*) the DC transformer and the trading strategies are applied to the full price series, producing a recommendation matrix $\mathbf{R} \in \{\text{Buy}, \text{Hold}, \text{Sell}\}^{n \times 70}$; (*iv*) GA and CMA-ES optimize the weight vector on the training portion of $\mathbf{R}$; and (*v*) the approach applies the optimized weights to the test portion of $\mathbf{R}$ without modification.

5.2 Data and Optimizer Configuration

We collected daily adjusted closing prices for 30 equities listed on B3, spanning sectors including banking, utilities, commodities, energy, retail, construction, and technology. Prices span January 2015 to December 2025 ($\sim$2 700 trading days per asset), and we obtained them using the `yfinance` Python package [21]. The framework runs on Python 3.11. As the risk-free rate, we use the daily Interbank Deposit Certificate (CDI), sourced from the Brazilian Central Bank time series system (series 12) [22]. Transaction costs are set at 3 basis points per side (0.03%), corresponding to the settlement and trading fees for retail equity swing trades [20].

The GA follows the configuration of Salman et al. [6] for direct comparability with the original MSTGAM baseline, with population size 150, 18 generations, two-point crossover with probability 0.95, tournament selection of size 2, and elitism of 1. The first 70 individuals are identity chromosomes, each assigning full weight to a single sub-strategy; the approach initialized the remaining 80 with random, normalized weights. CMA-ES operates in an unconstrained 70-dimensional space with initial step size $\sigma_0 = 0.3$, population size 50, and a maximum of 10 000 function evaluations. The population size exceeds Hansen's [8]

default of $4 + \lfloor 3 \ln 70 \rfloor = 16$ to increase exploration in the high-dimensional weight space.

Salman [6] executed the GA 50 times per configuration and retains the single best chromosome, effectively selecting the most favorable outcome across runs, as the walk-forward protocol already requires 270 independent optimizations across 30 tickers and 9 folds. Both optimizers use a fixed random seed per fold to ensure reproducibility. Each ticker completes in approximately 80 min on an Intel Core i7-10700 workstation (8 cores/16 threads running at 2.9–4.8 GHz) with 16 GB DDR4-2666 RAM, with individual folds ranging from 3 min (the shortest training window) to 15 min (the longest).

5.3 Evaluation Metrics

We evaluated each out-of-sample fold using four metrics. The primary metric is the annualized Sharpe Ratio (Eq. 2), which measures excess return over the CDI per unit of risk. Since the Brazilian risk-free rate has averaged approximately 10% per year over the sample period, positive Sharpe Ratios are less attainable than in lower-rate markets. The remaining metrics are total return, annualized volatility, and maximum drawdown (MDD). When the optimizer produces a weight vector that does not trigger trades, the portfolio remains completely in CDI throughout the fold. In this situation, the excess returns are near-zero with negligible variance, which can produce artificially high Sharpe Ratios. Empirical results from one GA fold showed that the absence of trades resulted in a Sharpe Ratio exceeding 200. To handle such cases, the Sharpe Ratio is substituted by zero in the corresponding metric whenever the standard deviation and the mean of excess returns are close to zero.

6 Results

This section presents the findings, organized by the four tested hypotheses: cross-asset performance (H1) in Sect. 6.1, risk reduction (H2) in Sect. 6.2, training window effect (H3) in Sect. 6.3, and optimizer generalization (H4) in Sect. 6.4.

6.1 Cross-Asset Performance (H1)

Table 2 summarizes the aggregate out-of-sample metrics. Both optimizers outperform Buy-and-Hold in risk-adjusted terms, with CMA-ES achieving a mean Sharpe Ratio of 0.868 compared to 0.637 for the passive benchmark. The profit factor nearly doubles that of Buy-and-Hold (8.47 vs. 4.43), indicating that the ensemble generates substantially larger gains in relation to its losses. The advantage holds across 26 of 30 tickers. Among the four in which Buy-and-Hold dominates (BBAS3, PETR4, SBSP3, SUZB3), three fall within commodity-linked sectors, where sustained directional trends over the sample period favor uninterrupted exposure. The per-ticker mean Sharpe Ratio for CMA-ES ranges from

Table 2. Aggregate out-of-sample metrics across 30 equities and 270 folds (2017–2025). Values in parentheses denote standard deviations across folds.

Metric	CMA-ES	GA	B&H
Mean Sharpe Ratio	0.87 (0.93)	0.85 (1.05)	0.64 (0.96)
Folds > B&H Sharpe	60%	59%	–
Tickers > B&H Sharpe	87%	83%	–
Profit factor ($\sum r^+ / \sum \lvert r^- \rvert$)	8.47	6.26	4.43
Vol. ratio ($\sigma_{ens}/\sigma_{B\&H}$)	0.68	0.64	1.00
Folds with lower MDD	95%	95%	–
MDD ratio ($MDD_{ens}/MDD_{B\&H}$)	0.68	0.65	1.00
Mean degradation ($SR_{train} - SR_{test}$)	0.54 (0.99)	1.08 (0.96)	–

0.12 (SUZB3) to 1.58 (VIVT3), with an interquartile range of 0.60–1.15, which suggests the aggregate result is robust to outliers.

Figure 1 illustrates this for BOVA11, the Ibovespa ETF. Over nine concatenated out-of-sample years, CMA-ES triples the initial capital while Buy-and-Hold roughly doubles it. The gap opens during the 2020 and 2024 corrections, where the ensemble moves to CDI and avoids the worst of the drawdown.

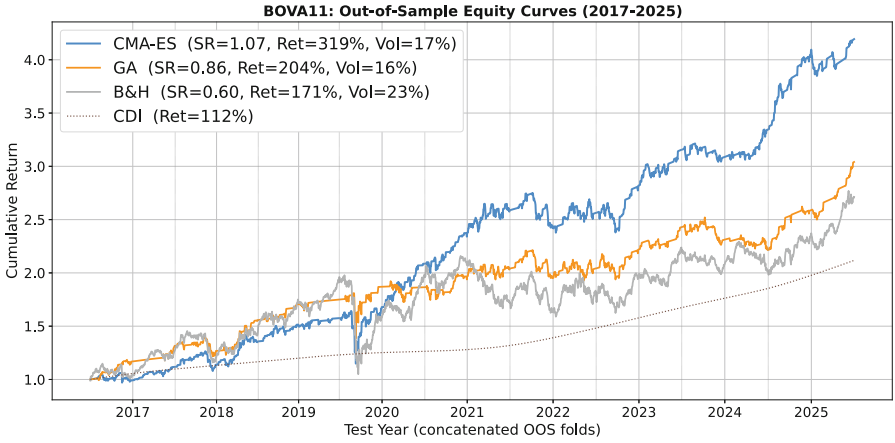


Fig. 1. BOVA11 out-of-sample equity curves (2017–2025), concatenated across nine folds. Legend reports cumulative Sharpe Ratio, return, and annualized volatility.

6.2 Risk Reduction (H2)

Figure 2 shows the annualized volatility of each fold under CMA-ES and GA against Buy-and-Hold. All folds lie below the diagonal for both optimizers: the

ensemble never increases volatility. The mean ratio is 0.68, a reduction of approximately 32%. Maximum drawdown follows the same pattern, with 95% of folds showing lower drawdowns under CMA-ES and GA.

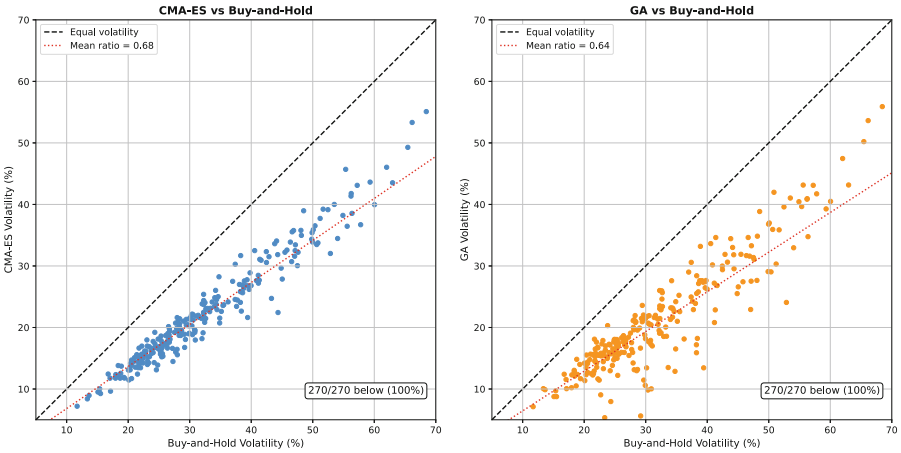


Fig. 2. Annualized volatility per fold: CMA-ES (left) and GA (right) versus Buy-and-Hold. All 270 points lie below the diagonal.

The mechanism behind this result is the ensemble's tendency to allocate the portfolio to CDI whenever the sub-strategies disagree. This filtering removes the asset's return during periods of ambiguous signal. Volatility reduction is therefore universal across tickers and folds.

6.3 Training Window Effect (H3)

Figure 3 shows the Sharpe difference between CMA-ES and Buy-and-Hold as a function of the training window length. Folds trained on shorter windows (2–5 years, $n = 120$) show a mean advantage of +0.10, while folds trained on longer windows (6–10 years, $n = 150$) show +0.33. The positive trend suggests that exposure to more market regimes, including the 2020 crisis, improves both strategy calibration and ensemble weight optimization.

6.4 Optimizer Generalization (H4)

CMA-ES and GA produce indistinguishable out-of-sample Sharpe Ratios (0.868 vs. 0.852), but differ in how they reach this result. Figure 4 shows the distribution of train-to-test Sharpe degradation for both methods. CMA-ES degrades by 0.54 on average (std = 0.99), while the GA degrades by 1.08 (std = 0.96), approximately twice as much. The GA achieves higher training Sharpe Ratios but fails to transfer this advantage to unseen data. While the GA explores the

search space through population-based selection and recombination without an explicit model of the underlying distribution, CMA-ES iteratively estimates and refines a full covariance matrix that parametrizes the search distribution itself. Thus, CMA-ES explores the 70-dimensional weight space more efficiently and converges to solutions that generalize better.

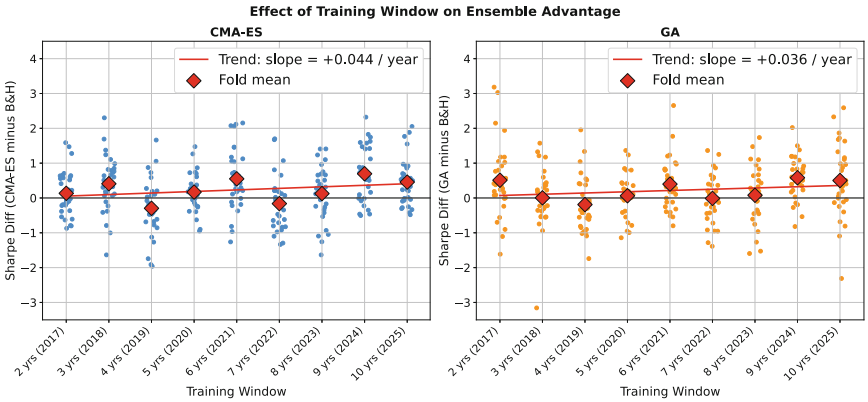


Fig. 3. Sharpe difference (CMA-ES minus B&H) versus training window length. Red diamonds indicate fold means; the red line is the linear trend. (Color figure online)

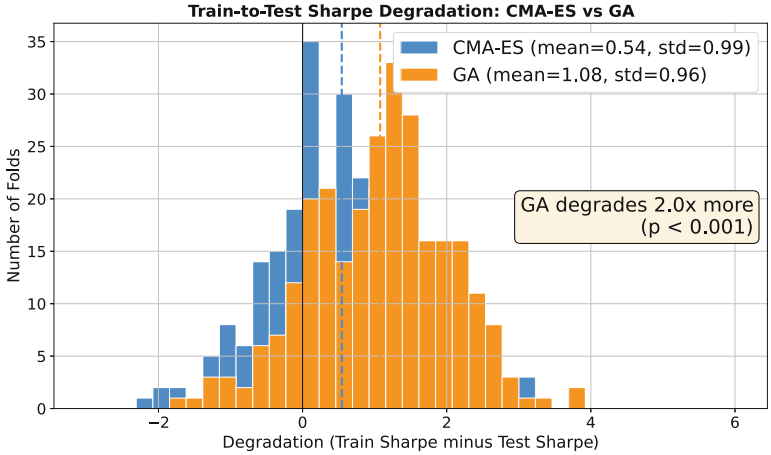


Fig. 4. Train-to-test Sharpe degradation across 270 folds. CMA-ES (blue) concentrates near zero; GA (orange) shifts toward higher degradation. (Color figure online)

7 Conclusion

We extended the MSTGAM framework for the Brazilian market by replacing its fixed thresholds with volatility-adaptive ones, introducing walk-forward validation, and adding CMA-ES as a second optimizer alongside the original GA. The results confirm the viability of DC-based trading in a high-interest-rate emerging market. The CMA-ES ensemble outperforms Buy-and-Hold in risk-adjusted terms (Sharpe 0.87 vs. 0.64, H1), with volatility reduced in every fold and drawdown in 95% of them (H2). Longer training windows strengthen this advantage (H3), suggesting the ensemble benefits from exposure to a wider range of market conditions. CMA-ES and GA deliver similar out-of-sample Sharpe Ratios, but CMA-ES degrades half as much from training to test (H4), making it the more reliable optimizer for this 70-dimensional weight space.

The most relevant limitation is the annual rebalancing cycle. Markets shift regimes within a calendar year, and the current protocol cannot react until the next fold boundary. A related issue is that the ensemble relies exclusively on DC-derived signals: it has no direct representation of the prevailing market regime and cannot reliably discern whether the market is trending or range-bound. Incorporating strategies less correlated with DC, such as classical TA indicators, could diversify the signal base and improve robustness during regimes where DC events are sparse.

On the deployment side, the framework offers structural advantages that simplify operation. Daily data refresh and per-fold recalibration are computationally inexpensive, allowing the model to incorporate new observations and retrain quickly between trading sessions. The modular structure of the ensemble accommodates new strategies with minimal changes: any rule that emits Buy, Hold, or Sell signals can be added as a sub-strategy, enabling the optimizer to assign its weight. These properties make the framework suitable for real-time operation on equities, where end-of-day decisions can be computed and queued for the closing auction.

The most promising direction for future work is regime-conditional weight optimization. We expect that certain DC strategies perform well in specific market conditions but poorly in others. Providing the optimizer with regime information as an additional input, rather than relying on a single static weight vector per fold, could allow the ensemble to alternate between strategy configurations as conditions change.

Acknowledgments. The second author acknowledges the Brazilian funding agency CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for financial support with grant 402599/2023-3. The authors also thank the anonymous referees for their careful reading, insightful comments, and suggestions.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Guillaume, D.M., Dacorogna, M.M., Davé, R.R., Müller, U.A., Olsen, R.B., Pictet, O.V.: From the bird's eye to the microscope: a survey of new stylized facts of the intra-daily foreign exchange markets. Finance Stochast. **1**(2), 95–129 (1997)
2. Tsang, E.P.K., Tao, R., Serguieva, A., Ma, S.: Profiling high-frequency equity price movements in directional changes. Quant. Finan. **17**(2), 217–229 (2017)
3. Aloud, M., Tsang, E.P.K., Olsen, R., Dupuis, A.: A directional-change event approach for studying financial time series. Economics **6**(1), 1–17 (2012)
4. Glattfelder, J.B., Dupuis, A., Olsen, R.B.: Patterns in high-frequency FX data: discovery of 12 empirical scaling laws. Quant. Finan. **11**(4), 599–614 (2011)
5. Salman, O., Melissourgos, T., Kampouridis, M.: A genetic algorithm for the optimization of multi-threshold trading strategies in the directional changes paradigm. Artif. Intell. Rev. **59**(1), 2 (2026)
6. Salman, O.: Trading Strategies Optimization by Genetic Algorithm Under the Directional Changes Paradigm. PhD Thesis, University of Essex (2024)
7. Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. Evol. Comput. **9**(2), 159–195 (2001)
8. Hansen, N.: The CMA Evolution Strategy: A Tutorial. arXiv preprint arXiv:1604.00772 (2016)
9. Vermetten, D., de Nobel, J., Wang, H., Doerr, C., Bäck, T.: Comparison of high-dimensional Bayesian optimization algorithms on BBOB. ACM Trans. Evol. Learn. Optim. **4**(3), 1–31 (2024)
10. Sharpe, W.F.: Mutual fund performance. J. Bus. **39**(1), 119–138 (1966)
11. Tao, R.: Using Directional Change for Information Extraction in Financial Market Data. PhD Thesis, University of Essex (2018)
12. Salman, O., Kampouridis, M., Jarchi, D.: Trading strategies optimization by genetic algorithm under the directional changes paradigm. In: IEEE Congress on Evolutionary Computation (CEC), pp. 1–8 (2022)
13. Salman, O., Melissourgos, T., Kampouridis, M.: Optimization of trading strategies using a genetic algorithm under the directional changes paradigm with multiple thresholds. In: IEEE Congress on Evolutionary Computation (CEC), pp. 1–8 (2023)
14. Long, X., Kampouridis, M., Kanellopoulos, P.: An in-depth investigation of genetic programming under physical time and directional change frameworks for algorithmic trading. IEEE Access **13**, 151001–151015 (2025)
15. Long, X., Kampouridis, M., Papastylianou, T.: Multi-objective genetic programming-based algorithmic trading, using directional changes and a modified Sharpe ratio score for identifying optimal trading strategies. Artif. Intell. Rev. **59**(2), 39 (2025)
16. Rayment, G., Kampouridis, M.: High frequency trading with deep reinforcement learning agents under a directional changes sampling framework. In: 2023 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 387–394 (2023)
17. Rayment, G., Kampouridis, M.: Enhancing high-frequency trading with deep reinforcement learning using advanced positional awareness under a directional changes paradigm. In: 2024 International Conference on Machine Learning and Applications (ICMLA), pp. 63–70 (2024)
18. Chinraj, K.F., Anthonysamy, S.S.: A hybrid framework integrating mathematical genetic algorithm and multi-criteria decision making for robust rule-based stock market strategy optimization. Math. Comput. Sci. **6**(4), 89–100 (2025)

19. Yu, K., Cheng, C.: Dynamic Portfolio Optimization Using Evolution Strategy. Stanford University, CS229 Final Report (2017)
20. B3 – Brasil, Bolsa, Balcão: Tarifas de ações e fundos de investimento – mercado à vista. Available at: https://www.b3.com.br/pt_br/produtos-e-servicos/tarifas/listados-a-vista-e-derivativos/renda-variavel/tarifas-de-acoes-e-fundos-de-investimento/a-vista/. Accessed March 2026
21. Aroussi, R.: yfinance: Download market data from Yahoo! Finance API. Python package. https://pypi.org/project/yfinance/ (2017)
22. Banco Central do Brasil: Sistema Gerenciador de Séries Temporais (SGS), série 12 – Taxa de depósito interfinanceiro (CDI). https://www3.bcb.gov.br/sgspub/ (accessed March 2026)

Optimization and Technical-Economic Evaluation of DC Microgrids with Integrated Renewable Energy Sources

Adelaide Cerveira[1,2(✉)], Paulo Silva[3], and José Baptista[2,3]

[1] Department of Mathematics, University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal
[2] INESCTEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
{cerveira,baptista}@utad.pt
[3] Department of Engineering, University of Trás-os-Montes and Alto Douro, 5000-801 Vila Real, Portugal
al72010@alunos.utad.pt

Abstract. The growing use of renewable energy sources has made the development of efficient and sustainable microgrids increasingly important. In this context, direct current (DC) microgrids have emerged as a suitable solution, as they are more efficient and compatible with renewable energy generation and storage systems.

A mixed-integer linear programming (MILP) model is proposed to optimize the design of DC microgrids comprising photovoltaic modules, wind turbines and energy storage systems. The model determines the optimal number of energy storage systems and generation components. Furthermore, energy exchanges with the main grid are also determined over a long-term planning horizon.

The proposed model is evaluated using five case studies corresponding to different storage configurations and grid interaction scenarios. These case studies are based on seasonal load profiles derived from actual consumption data and solar irradiance data collected between 2019 and 2023. Technical and financial metrics, such as net present value (NPV), levelized cost of electricity (LCOE), internal rate of return (IRR), and payback period (PBP), are used to evaluate the system's performance.

Results show that photovoltaic-based configurations provide the most cost-effective solutions under current market conditions, while the inclusion of storage systems improves long-term profitability despite higher initial investment. Sensitivity analysis highlights the strong influence of energy prices and equipment costs on the optimal configuration.

Keywords: Direct Current Microgrids · Mixed Integer Linear Programming · Renewable Energy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 215–230, 2027.
https://doi.org/10.1007/978-3-032-30521-3_14

1 Introduction

With increasing ecological and economic concerns, several countries are providing incentives for the production of clean and sustainable energy. Renewable energy plays an important role in reducing greenhouse gas emissions while contributing to electricity production [21]. This is in line with the Paris Agreement [18], established in 2015, which represents an important milestone in the fight against climate change. The main objective of this agreement is to limit the rise in global temperatures and promote sustainable practices, such as the use of renewable resources.

The growing use of renewable energy sources (RES) leads to considerable challenges for the planning and operation of modern electricity systems. The variability and intermittency of sources such as solar and wind power require flexible and efficient solutions to ensure system reliability and economic viability.

Microgrids (MG) have emerged as an essential component in the development of smart grids. These are small-scale energy systems designed to maximize energy efficiency. They capitalize on production potential, enabling more efficient resource management and contributing to greater system flexibility and resilience. The importance of microgrids in promoting energy efficiency is highlighted by Directive 2018/2002 [5]. These decentralized systems for energy production and distribution facilitate the efficient integration of renewable sources. Flexible operational strategies and optimization algorithms provide microgrids with versatile and effective solutions to contemporary energy demands [20].

This paper presents an MILP model to determine the optimal size and operation of DC microgrids. The system includes photovoltaic and wind generation and considers long-term planning based on seasonal representative days. Different system configurations are analyzed, including the evaluation of techno-economic indicators such as Net Present Value (NPV), Levelized Cost of Energy (LCOE), Internal Rate of Return (IRR), and Payback Period (PBP). In addition, a sensitivity analysis is used to see how changes in energy prices and equipment costs affect system performance.

The paper is organized as follows. Section 2 briefly reviews the related literature. Section 3 presents the optimization model and describes techno-economic indicators. Section 4 describes the case study and Sect. 5 discusses the results. Finally, Sect. 6 presents the main conclusions of the work.

2 Microgrid Optimization: Literature Review

Microgrids represent an innovative advancement in the field of energy distribution, characterized by low-voltage networks that efficiently integrate diverse energy sources, loads, and storage systems [6,9]. Several works in the field of microgrids are referenced below. The article [1] provides a detailed analysis of microgrid architectures and their applications, exploring how MG, composed of both renewable and conventional energy sources, can improve the efficiency and sustainability of electrical systems. The study in [7] examines the transition between grid-connected and islanded operation modes of microgrids. The

main focus is to ensure a smooth transition between these modes without interruptions in energy supply. The work presented in [4] provides an overview of microgrids and practical guidelines for their implementation and operation, highlighting their numerous benefits, including increased resilience and reliability of the energy supply, particularly during main grid failures.

The integration of energy storage systems (ESS) is a promising solution to overcome some limitations and facilitate grid stability. The combination of ESS with renewable energies in microgrids can reduce reliance on conventional energy sources, enhance system resilience and reliability, and contribute to grid decarbonization [19]. Many countries are advancing toward the adoption of smart grids to achieve goals such as environmental sustainability, adequate energy supply, and the integration of electric vehicles, where RES and ESS play a crucial role [2]

To ensure the effective and sustainable operation of renewable energy systems, it is crucial to employ optimization techniques that capture their complexity. Among the widely used methods, Mixed-Integer Linear Programming (MILP) stands out for its ability to model both discrete and continuous variables, providing a powerful framework for addressing the complex decisions involved in integrating, managing, and optimizing renewable energy sources and storage systems. This versatility and effectiveness have made MILP a widely adopted approach in the literature for optimizing renewable energy systems.

In [12,17], the authors employed an MILP model to optimize hybrid renewable energy systems to meet winery demand, accounting for equipment installation costs and environmental constraints related to sustainability.

The article [13] uses a MILP model to optimize the management of an isolated microgrid with storage, accounting for uncertainties in load and renewable energy production. The main objective is to minimize operating costs while ensuring the reliability and continuity of the energy supply.

Although much work has been done in this field, significant challenges remain. It is important to undertake long-term planning, account for component deterioration, and combine optimization techniques with the assessment of technical and economic indicators.

In this context, this paper proposes a MILP-based optimization model for the design and operation of a DC microgrid integrating renewable energy sources and storage systems. The proposed approach incorporates long-term planning, seasonal variability, and a comprehensive techno-economic assessment, providing a more integrated framework for microgrid analysis and decision-making.

3 Optimization Model and Techno-Economic Indicators

This section presents the optimization model and techno-economic indicators.

3.1 Optimization Model

The aim of this work is to design microgrids with renewable energy sources in order to maximize overall benefits. The grid integrates solar and wind power

generation and is supported by energy storage systems to efficiently manage charging and discharging operations, thereby minimizing energy purchases over a 20-year horizon. The problem is formulated as a MILP model to optimize energy production, storage, and distribution, whilst simultaneously respecting technical and economic constraints. The model formulation considers four typical days for each season, with each day divided into 24 h periods, enabling a detailed analysis of the system behavior throughout the year.

Sets and Parameters.

- $D = \{1, \ldots, nd\}$, set of $nd = 4$ periods of a day, one for each season of the year;
- $H = \{1, \ldots, nh\}$, set of $nh = 24$ hourly periods, corresponding to one day;
- $A = \{1, \ldots, na\}$, set of $na = 20$ periods of a year, which is the project lifespan;
- k^{bat}, battery capacity, in kWh;
- pc_{ij}, energy purchase price, in €/kWh, in hour $i \in H$ and day $j \in D$;
- pv_{ij}, energy sale price, in €/kWh, in hour $i \in H$ and day $j \in D$;
- dc_{ij}, energy consumption, in kWh, in hour $i \in H$ and day $j \in D$;
- epv_{ij}, solar energy, in kWh, produced by a solar panel in hour $i \in H$ and day $j \in D$;
- et_{ij}, wind energy, in kWh, produced by a wind turbine in hour $i \in H$ and day $j \in D$;
- cm^{pv}, annual maintenance cost, in €, for the solar panel;
- cm^{t}, annual maintenance cost, in €, for the wind turbine;
- cm^{bat}, annual maintenance cost, in €, for the battery;
- cm^{inv}, annual maintenance cost, in €, for the inverter;
- ca^{pv}, acquisition cost, in €, of a solar panel;
- ca^{t}, acquisition cost, in €, of a wind turbine;
- ca^{inv}, acquisition cost, in €, of an inverter;
- ca^{bat}, acquisition cost, in €, of a battery;

Decision Variables.

- x^{pv}, integer variable representing the number of solar panels to be installed;
- x^{t}, integer variable representing the number of wind turbines to be installed;
- yc_{ija}, energy, in kWh, bought from the grid during hour $i \in H$, day $j \in D$, and year $a \in A$;
- yv_{ija}, energy, in kWh, sold to the grid during the hour $i \in H$, day $j \in D$, and year $a \in A$;
- z^{bat} binary variable that takes the value 1 if batteries are used;
- soc_{ija}, battery state of charge at the beginning of hour $i \in H$ (with $i = 1$ representing the initial state), day $j \in D$, and year $a \in A$;
- yib_{ija}, energy, in kWh, injected into the battery in hour $i \in H$, day $j \in D$, and year $a \in A$;
- ydb_{ija}, energy, in kWh, discharged from the battery in hour $i \in H$, day $j \in D$, and year $a \in A$;

- wib_{ija}, binary variable that takes the value 1 if there is energy injection into the battery and 0 otherwise, in hour $i \in H$, day $j \in D$, and year $a \in A$;
- wdb_{ija}, binary variable that takes the value 1 if the battery is discharging and 0 otherwise, in hour $i \in H$, day $j \in D$, and year $a \in A$;
- wc_{ija}, binary variable that takes the value 1 if energy is bought and 0 otherwise, in hour $i \in H$, day $j \in D$, and year $a \in A$;
- wv_{ija}, binary variable that takes the value 1 if energy is sold and 0 otherwise, in hour $i \in H$, day $j \in D$, and year $a \in A$;
- wb_{ija}, binary variable that takes the value 1 if batteries are used and 0 otherwise, in hour $i \in H$, day $j \in D$, and year $a \in A$;

Objective Function. The objective function (1) corresponds to the maximization of the total profit, TP, over the 20-year planning horizon. This profit is obtained by subtracting the total costs, which include energy purchase costs, equipment purchase costs, and maintenance expenses, from the revenue generated by energy sales. Additionally, an inflation rate of 1.39% is considered over the 20 years.

$$\begin{aligned} \max \quad TP = 91.25 \cdot \sum_{i \in H} \sum_{j \in D} \sum_{a \in A} \left(yv_{ija} \cdot pv_{ij} - yc_{ija} \cdot pc_{ij} \right) \cdot (1.0139)^{(a-1)} \\ - x^{pv} \cdot ca^{pv} - x^{t} \cdot ca^{t} - z^{bat} \cdot ca^{bat} - z^{bat} \cdot ca^{bat} \cdot (1.0139)^{10} - ca^{inv} \\ - \sum_{a \in A} \left(x^{pv} \cdot cm^{pv} + x^{t} \cdot cm^{t} + z^{bat} \cdot cm^{bat} + cm^{inv} \right) \cdot (1.0139)^{(a-1)} \end{aligned} \tag{1}$$

As mentioned earlier, a representative day was selected for each season of the year. To extrapolate the results to the entire year, the corresponding value was multiplied by 91.25, which represents the average number of days per season. Battery investment is modeled considering a unit lifetime of 10 years. Therefore, when battery storage is selected, an initial investment is incurred in the first year and a replacement cost is considered in year 11 of the project horizon.

Constraints. Constraint (2) enforces the energy balance for each hour $i \in H$, day $j \in D$, and year $a \in A$:

$$\begin{aligned} \forall i \in H, j \in D, a \in A, \quad x^{pv} \cdot epv_{ij} \cdot (1 - 0.0055)^{(a-1)} + x^{t} \cdot et_{ij} \cdot (1 - 0.015)^{(a-1)} \\ + ydb_{ija} + yc_{ija} = dc_{ij} + yib_{ija} + yv_{ija} \end{aligned} \tag{2}$$

The left-hand side represents the available energy (generation, battery discharge, and grid import), while the right-hand side accounts for demand and energy allocation (battery charge and energy sold). The parameters 0.0055 (0.55%) and 0.015 (1.5%) represent the annual degradation rates of photovoltaic panels and wind turbines, respectively.

The state of charge (SOC) of the batteries at the beginning and at the end of each day is defined by constraints (3) and (4):

$$\forall j \in D, a \in A, \quad soc_{1ja} = 0.0 \tag{3}$$

$$\forall j \in D, a \in A, \quad soc_{25,ja} = 0.0 \tag{4}$$

Here, $soc_{25,ja}$ represents the state of charge at the end of the representative day (start of hour 25), which corresponds to the end of hour 24, ensuring that battery operation can be consistently replicated across days of the same season.

The SOC of the batteries must also satisfy the following limits:

$$\forall i \in H, j \in D, a \in A, \quad soc_{ija} \geq 0.0 \tag{5}$$

$$\forall i \in H, j \in D, a \in A, \quad soc_{ija} \leq 0.9 \cdot k^{bat} \cdot z^{bat} \tag{6}$$

The dynamics of the SOC between consecutive hours is given by:

$$\forall i \in H, j \in D, a \in A, \quad soc_{i+1,ja} - soc_{ija} = yib_{ija} - ydb_{ija} \tag{7}$$

The energy injected into and discharged from the battery is constrained by the binary variables and the battery capacity:

$$\forall i \in H, j \in D, a \in A, \quad \begin{cases} yib_{ija} \leq 0.9 \cdot k^{bat} \cdot wib_{ija} \\ yib_{ija} \leq 0.9 \cdot k^{bat} \cdot z^{bat} \\ ydb_{ija} \leq 0.9 \cdot k^{bat} \cdot wdb_{ija} \\ ydb_{ija} \leq 0.9 \cdot k^{bat} \cdot z^{bat} \end{cases} \tag{8}$$

The binary variables ensure that charging and discharging cannot occur simultaneously:

$$\forall i \in H, j \in D, a \in A, \quad wib_{ija} + wdb_{ija} \leq 1 \tag{9}$$

Finally, an upper limit on the energy charged or discharged per hour is imposed, according to the battery's rated power.

$$\forall i \in H, j \in D, a \in A, \quad yib_{ija} \leq 5.0, \quad ydb_{ija} \leq 5.0 \tag{10}$$

These constraints ensure that the battery, when installed, operates physically consistently and within its technical limits, while avoiding simultaneous charging and discharging.

The quantities of energy purchased (yc_{ija}) and sold (yv_{ija}) are continuous decision variables representing, respectively, the amount of energy bought from and sold to the grid during hour $i \in H$, day $j \in D$, and year $a \in A$. Their operation is controlled by the binary variables wc_{ija} and wv_{ija}, which indicate whether a purchase or sale occurs in a given period.

The following linear constraints then ensure a proper logical relationship between the continuous energy variables and the corresponding binary variables:

$$\forall i \in H, j \in D, a \in A, \quad yv_{ija} \leq wv_{ija} \cdot M, \quad yc_{ija} \leq wc_{ija} \cdot M \tag{11}$$

where M is a large constant, representing an upper bound on the energy that could be purchased or sold in a single period.

In each period, the purchase and sale of energy cannot occur simultaneously,

$$\forall i \in H, j \in D, a \in A, \quad wv_{ija} + wc_{ija} \leq 1 \tag{12}$$

The constraint (13) represents the limitation on the number of photovoltaic modules that can be installed, due to the physical space available on the roof.

$$x^{pv} \leq 48 \tag{13}$$

Constraint (14) imposes non-negativity on the injected and discharged energy from the battery.

$$\forall i \in H, j \in D, a \in A, \quad yib_{ija} \geq 0, ydb_{ija} \geq 0 \tag{14}$$

Constraint (15) imposes non-negativity on the number of installed turbines.

$$x^{t} \geq 0 \tag{15}$$

Constraints (16) specify the binary nature of a subset of the decision variables.

$$\forall i \in H, j \in D, a \in A, \quad wib_{ija}, wdb_{ija}, wc_{ija}, wv_{ija}, wb_{ija}, z^{bat} \in \{0, 1\} \tag{16}$$

Changes in parameters, such as the cost of the components, renewable energy production, and load profiles, have a substantial impact on the results. They alter the cost-benefit dynamics and modify the microgrid's operational strategies. These parameters can vary significantly between regions (such as solar radiation, temperature, and wind speed), affecting the technical and economic performance and the viability of the microgrid. The model presented is easily adaptable to other data, facilitating its application in different locations without requiring major structural modifications.

3.2 Techno-Economic Indicators

To evaluate the economic performance of the proposed microgrid configurations, several standard indicators are considered, including LCOE, NPV, IRR, and PBP [3].

The LCOE (17) represents the average cost of energy production over the system lifetime, accounting for investment, operation, and maintenance costs. The NPV (18) evaluates the profitability of the investment by comparing discounted cash inflows and outflows. The IRR indicates the expected return rate

of the project, while the PBP measures the time required to recover the initial investment, (19) and (20).

$$\text{LCOE} = \frac{\sum_{n=1}^{N} \frac{C_n + O_n + M_n}{(1+i)^n}}{\sum_{n=1}^{N} \frac{Q_n}{(1+i)^n}} \tag{17}$$

$$NPV = \sum_{n=1}^{N} \frac{R_{ln}}{(1+i)^n} - I_t \tag{18}$$

$$\sum_{n=1}^{N} \frac{R_{ln}}{(1+IRR)^n} - I_t = 0 \tag{19}$$

$$PBP = \frac{\sum_{n=1}^{N} \frac{R_{ln}}{(1+i)^n}}{I_t} \tag{20}$$

where C_n represents the capital costs or initial investment, O_n denotes the operational costs, M_n refers to the maintenance costs, Q_n is the energy produced in year n, i is the discount rate, N represents the project lifetime, in years, where R_{ln} represents the net revenue of year n, and I_t represents the total investment updated to year 0.

These indicators are computed based on the optimized system configuration and are used to compare the economic viability of the different case studies.

4 Case Study: Microgrid Description

This section presents the case study used to validate the proposed optimization model. The analyzed system is a grid-connected microgrid integrating renewable energy sources, storage, and a consumer load, enabling the assessment of operational and economic performance under realistic conditions.

The microgrid consists of a photovoltaic (PV) system, a wind turbine, an energy storage system (ESS), a consumer load, and a bidirectional DC/AC inverter, as illustrated in Fig. 1. The system is connected to the main grid, allowing both energy import and export.

The PV generation is modeled using a JAM66S30-505/MR/1500V module [16], considering typical performance parameters. Solar irradiance and temperature data were obtained from the Photovoltaic Geographical Information System (PVGIS) platform, using average daily profiles for representative months (January, April, July, and November) over the period 20192023, for the Porto region, with a tilt angle of 35°.

Wind generation is obtained using a Aeolos-V 10 kW turbine model [15]. Wind speed data were obtained from the Portuguese Institute for Sea and Atmosphere (IPMA), based on historical measurements from a meteorological station in Porto.

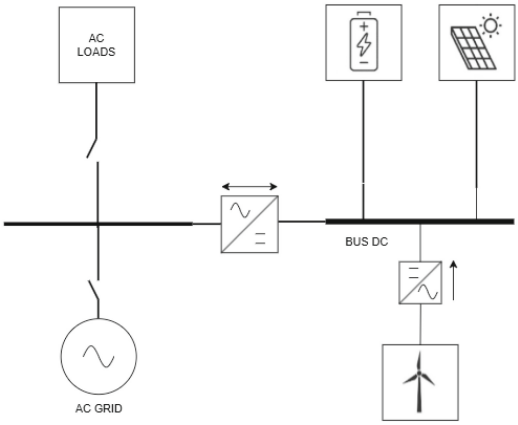


Fig. 1. DC Microgrid block diagram

The storage system is modeled as a lithium-ion battery, Huawei LUNA2000-10-SO [11], with typical operational characteristics, including charging/discharging limits and efficiency. A nominal capacity of 10 kWh is considered, with constraints on maximum charging and discharging power.

DC/AC inverters play a crucial role in renewable energy production systems. They are responsible for converting the energy produced by photovoltaic modules and wind turbines, which is in direct current (DC), into alternating current (AC), which is used by electrical equipment and the power grid. The FoxESS 25kW inverter [10] stands out as a modern and reliable solution for medium- and large-scale applications. This model is bidirectional, enabling operation in two energy flow directions. It converts renewable energy or energy discharged from the battery to power the electrical grid or local consumption. Additionally, when energy prices are low, the system can acquire energy from the grid, converting the alternating current from the grid into direct current to recharge the batteries for later use. This bidirectional flow optimizes the use of renewable energy and reduces energy costs.

The data for the load profiles were obtained from the E-REDES Open Data platform. Four representative days, each corresponding to a different season of the year, for consumers from the Porto region were considered, allowing the analysis of seasonal variability in demand.

The energy purchase and sale prices were obtained from the Iberian Electricity Market Operator [14], based on the average of 24 h of each day in the months of January, April, July, and November over the last five years.

The values considered for the cost of purchasing and maintaining energy production and storage components are based on market data.

5 Results and Discussion

This study analyzes the feasibility and operational efficiency of different microgrid configurations and compares their performance across various operational scenarios. Five main case studies were developed, Case Study I, II, III, IV and V. Each one represents a distinct combination of renewable energy generation, energy storage systems, and grid interactions, providing a solid basis for technical and economic comparisons. The proposed optimization model determines the optimal number of PV modules and wind turbines, as well as the number of batteries, and manages energy exchange. Hourly load profiles, solar irradiance, and wind speed data were incorporated to capture on-site operating conditions. The model aims to balance production with consumption while minimizing costs, accounting for constraints such as physical installation limits, production capacity, and the operational requirements of the energy storage system (ESS).

Case Study I serves as a baseline for calculating savings in other cases. In this case, the system relies entirely on the electrical grid, with no PV panels, wind turbines, or batteries. The total energy acquisition cost to meet the load is 119,472.59 €. Therefore, in the other cases, we can compute the savings by comparing with this value.

Case Study II does not include the use of storage systems and considers the possibility of installing photovoltaic panels and wind turbines. In the optimal solution, 24 PV modules are installed and no wind turbines, resulting in equipment and maintenance costs of 3,767.18 € and 1,292.61 €, respectively. These amounts represent a total savings of 20,244.62 € when compared to the baseline scenario.

Case Study III still does not consider the use of storage systems, but, in order to obtain a hybrid system, at least one wind turbine is required. The model selects 22 PV modules and 1 wind turbine, resulting in equipment cost of 23,388.68 € and maintenance cost of 8,025.20 €. Comparing with the baseline scenario, this system yields a negative net benefit of 545.46 €, indicating that this configuration is not economically viable.

In Case Study IV, two 10-kWh battery systems are installed, enabling energy storage to optimize system operation and grid interaction. In the optimal solution, 48 PV modules are installed, with no wind turbines. Equipment and maintenance costs are 16,684.38 € and 5,724.80 €, respectively. In this case, there is a savings of 30,589.75 € compared to the baseline, highlighting the storage system's significant impact on its long-term profitability. An additional case study (Case V) was also considered, which includes the use of batteries and requires the installation of at least one wind turbine. This case did not demonstrate economic advantages compared to the previous cases.

Techno-Economic Analysis

In Fig. 2 the LCOE for each case study is presented. This indicator is used to evaluate the average energy production cost over the system's lifetime. A discount rate of 4% was used, obtained from [8]. Case II exhibits the lowest LCOE

due to the adoption of low-cost renewable technology, whereas configurations incorporating batteries or wind turbines incur higher LCOE values.

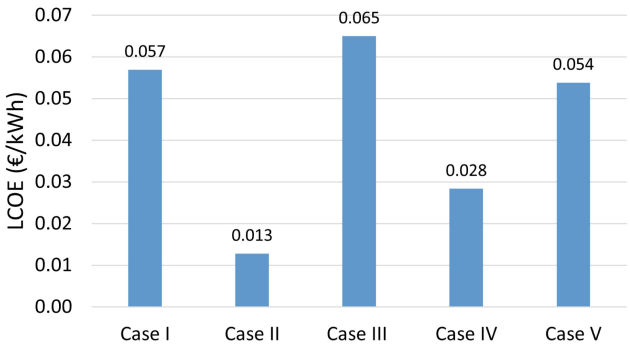


Fig. 2. Levelized Cost of Energy for all the cases studies

Figure 3 presents the NPV for all cases. The obtained values indicate that Cases II and IV are profitable, while Cases III and V are economically unattractive.

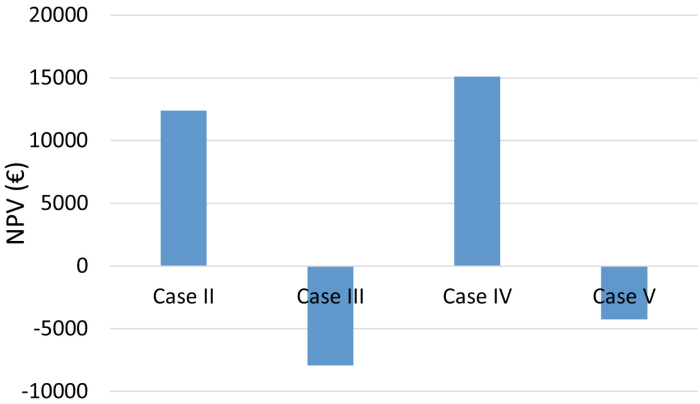


Fig. 3. Net Present Value for different cases studies

The IRR and PBP assessments are presented in Table 1. It can be concluded that Case II achieves the highest IRR (30.13%) and the shortest PBP (3.14 years), while Case IV provides moderate returns (IRR 12.43%, PBP 7.07 years) with substantial cumulative profit. Cases III and V exhibit low or negative IRR and a prolonged PBP, which underscores the limited financial viability of hybrid systems that include wind turbines at the chosen site for the system installation.

Table 1. Internal Rate of Return and Payback Period for each case

Case	IRR (%)	PBP (years)
Case II	30.13%	3.14
Case III	-0.22%	20.48
Case IV	12.43%	7.07
Case V	2.67%	15.27

Fig. 4 illustrates the amortization over the project's lifetime, showing how the invested value is progressively recovered throughout the system's lifespan.

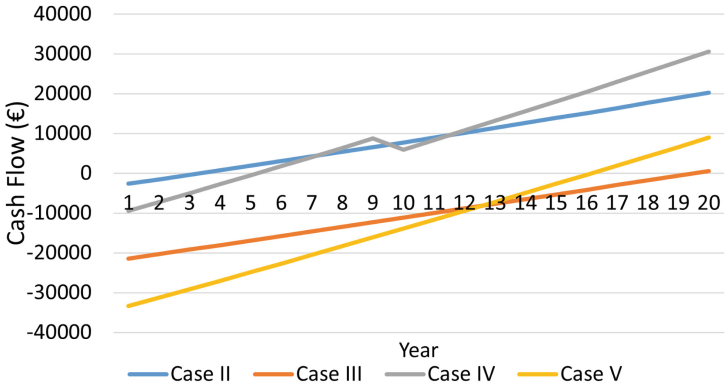


Fig. 4. Cash Flow during project life time for different cases studies

Sensitivity Analysis

To assess the robustness of the system under market fluctuations, Case IV was re-evaluated under variations in energy prices and equipment costs. The analysis included a ±20% change in purchase and sale prices, a 20% reduction in PV and wind turbine costs, and a 40% reduction in wind turbine costs. Results indicate that higher energy prices (Case VI) increase profitability, achieving the highest IRR (15.84%) and shortest PBP (5.77 years). Reductions in equipment costs (Cases VIII and IX) moderately improve economic performance, with Case IX (40% reduction in wind turbine cost) yielding the lowest LCOE and enhanced cost-effectiveness (Tables 2 and 3, Figs. 5 and 6).

Discussion

Overall, the results demonstrate that PV-based configurations are the most cost-effective and reliable solution in the short-to-medium term, while storage integration enhances long-term profitability and operational flexibility. Wind turbines,

Table 2. Summary of sensitivity analysis results

Case	Optimal configuration	Savings (€)
Case VI (+20% purchase & sale price)	48 PV + 2 Bat	41188.39
Case VII (-20% purchase & sale price)	48 PV + 2 Bat	19989.39
Case VIII (-20% PV & Turbine cost)	48 PV + 2 Bat	32050.68
Case IX (-40% Turbine cost)	48 PV + 2 Bat + 4 Turb	34031.32

Table 3. Internal Rate of Return and Payback Period for the new case studies

Case	IRR (%)	PBP (years)
Case VI	15.84%	5.77
Case VII	8.74%	9.10
Case VIII	13.64%	6.55
Case IX	4.42%	13.08

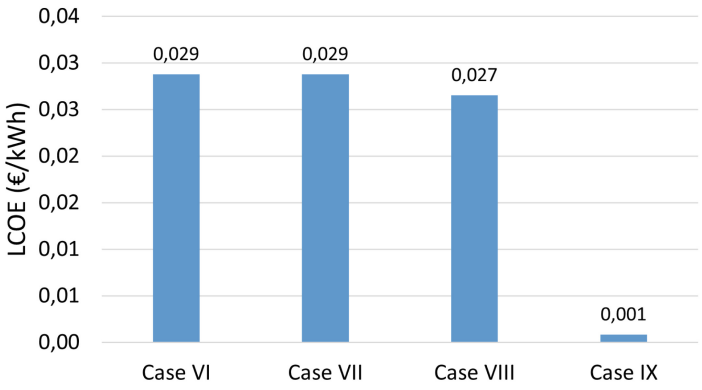


Fig. 5. Net Present Value for different cases studies

despite offering energy diversification, contribute less to economic feasibility due to higher initial costs and suboptimal performance in the studied context. Sensitivity analysis confirms the strong dependence of microgrid profitability on energy market prices and, to a lesser extent, on equipment cost reductions. The proposed model is adaptable for future integration of alternative storage technologies, such as thermal or hydrogen systems, which could further enhance flexibility and sustainability in hybrid microgrid designs.

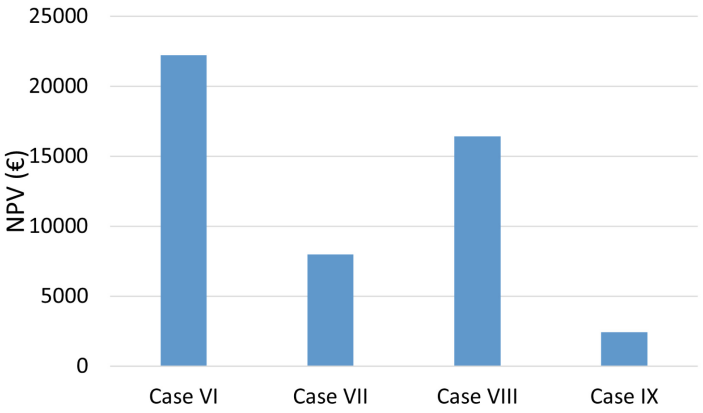


Fig. 6. Net Present Value for sensitivity analysis case studies (VIIX)

6 Conclusions

This study analyzed five microgrid configurations to evaluate the impact of renewable energy integration and energy storage systems on operational efficiency and economic performance. The main objective was to identify the most effective system configuration to maximize operational profit, reduce dependence on the electrical grid, and improve sustainability.

Case II, with the installation of 24 photovoltaic modules without storage, achieved the best performance in terms of cost-effectiveness, Internal Rate of Return (30.13%), and shortest Payback Period (3.14 years). So, the PV-based solutions are highly efficient and economically viable for the analyzed scenario. In contrast, Case IV, which incorporates 48 PV modules and two batteries, provided the highest cumulative profit over a 20-year horizon (30,589.75€), highlighting the long-term advantages of combining PV energy production with storage capacity.

The inclusion of batteries in Case IV enabled energy storage during periods of high solar production for later use during periods of low production, or to meet demand or sell it to the grid, thereby increasing the system's autonomy and operational flexibility. However, the higher initial investment associated with the batteries limited short-term economic performance. The integration of wind turbines did not prove beneficial in the context studied. Although it contributes to production diversification, the high installation cost of wind turbines and their lower-than-expected performance reduced economic viability compared to configurations dominated by photovoltaic energy.

The sensitivity analysis performed demonstrated the strong dependence of microgrid profitability on energy market prices. A 20% increase in purchase and sale prices (Case VI) significantly enhanced economic performance, achieving the highest IRR and shortest Payback Period, while a 20% decrease (Case VII) resulted in negative returns. Reductions in equipment costs (Cases VIII and

IX) provided moderate improvements, particularly through wind turbine cost reductions, which led to lower LCOE and improved techno-economic performance. Nevertheless, the effect of PV cost reductions is limited by the available installation area, which caps the number of panels at 48 and restricts additional benefits. These results indicate that, once spatial constraints are reached, further reductions in PV costs have diminishing returns, shifting the optimization focus toward other system components or design strategies.

The proposed optimization model is flexible and can accommodate alternative storage technologies, such as thermal or hydrogen systems, which could improve system flexibility and resilience in hybrid microgrid designs. Future research may explore these options to fully assess their long-term techno-economic benefits.

Overall, the results indicate that PV-based systems are the most attractive solution for short- to medium-term projects, while storage integration offers substantial advantages in long-term profitability and operational management, providing valuable guidance for microgrid design and investment decisions.

Acknowledgments. The first author's research was partially supported by FCT through project LA/P/0063/2020.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdelgawad, H., Sood, V.K.: A comprehensive review on microgrid architectures for distributed generation. In: 2019 IEEE Electrical Power and Energy Conference (EPEC), pp. 1–8 (2019). https://doi.org/10.1109/EPEC47565.2019.9074800
2. Babu, T.S., Vasudevan, K.R., Ramachandaramurthy, V.K., Sani, S.B., Chemud, S., Lajim, R.M.: A comprehensive review of hybrid energy storage systems: converter topologies, control strategies and future prospects. IEEE Access **8**, 148702–148721 (2020). https://doi.org/10.1109/ACCESS.2020.3015919
3. Baptista, J., Vargas, P., Ferreira, J.: A techno-economic analysis of floating photovoltaic systems, for southern European countries. Renew. Energy Power Qual. J. **19**, 57–62 (09 2021). https://doi.org/10.24084/repqj19.214
4. Cagnano, A., De Tuglie, E., Mancarella, P.: Microgrids: Overview and guidelines for practical implementations and operation. Appl. Energy **258**, 114039 (2020). https://doi.org/10.1016/j.apenergy.2019.114039, https://www.sciencedirect.com/science/article/pii/S030626191931726X
5. Council of European Union: Directive (eu) 2018/2002 - amending directive 2012/27/eu on energy efficiency (December 2018). https://eur-lex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:32018L2002. Accessed 15 Jan 2024
6. Dragičević, T., Lu, X., Vasquez, J.C., Guerrero, J.M.: Dc microgrids-part ii: a review of power architectures, applications, and standardization issues. IEEE Trans. Power Electron. **31**(5), 3528–3549 (2016). https://doi.org/10.1109/TPEL.2015.2464277

7. D'silva, S., Shadmand, M., Bayhan, S., Abu-Rub, H.: Towards grid of microgrids: Seamless transition between grid-connected and islanded modes of operation. IEEE Open J. Industrial Electron. Society **1**, 66–81 (2020). https://doi.org/10.1109/OJIES.2020.2988618
8. European Commission: Communication from the commission on the revision of the method for setting the reference and discount rates. https://competition-policy.ec.europa.eu/state-aid/legislation/reference-discount-rates-and-recovery-interest-rates/reference-and-discount-rates_en. Accessed Aug 2024
9. Farrokhabadi, M.: Microgrid stability definitions, analysis, and examples. IEEE Trans. Power Syst. **35**(1), 13–29 (2020). https://doi.org/10.1109/TPWRS.2019.2925703
10. FoxESS: Foxess t25 g3 datasheet (2023). Accessed on Aug 2024
11. Huawei Technologies Co., L.: Luna2000 5/10/15-s0 storage system. https://solar.huawei.com/de/professionals/all-products/LUNA2000-5-10-15-S0/specs. Accessed Aug 2024
12. Jesus, B., Cerveira, A., Santos, E., Baptista, J.: The impact of optimizing hybrid renewable energy system on wine industry sustainability. In: 2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON), pp. 278–283 (June 2024). https://doi.org/10.1109/MELECON56669.2024.10608594
13. Li, Y., Yang, Z., Li, G., Zhao, D., Tian, W.: Optimal scheduling of an isolated microgrid with battery storage considering load and renewable generation uncertainties. IEEE Trans. Industr. Electron. **66**(2), 1565–1575 (2019). https://doi.org/10.1109/TIE.2018.2840498
14. Nominated electricity market operator: Omie. https://www.omie.es/. Accessed Aug 2024
15. Renugen: Aeolos v-10kw wind turbine brochure (2024). https://www.renugen.co.uk/content/small_wind_turbine_brochures/small_wind_turbine_brochures/Aeolos%20Wind%20Turbine/Aeolos-Aeolos-V-10kW-10kW-Wind-Turbine-Brochure.pdf. Accessed Aug 2024
16. Solar, J.: Ja solar technology: Performance and quality guarantees. https://www.jasolar.com/uploadfile/2021/0706/20210706053433945.pdf. Accessed Aug 2024
17. Teixeira, R., Cerveira, A., Silva, A., Baptista, J.: Hybrid renewable energy system optimisation for application in the winemaking sector. In: 2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON), pp. 272–277 (2024). https://doi.org/10.1109/MELECON56669.2024.10608472
18. United Nations Framework Convention on Climate Change: The paris agreement (2015). https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement
19. Yuqiang, W., Rong, Z., Yue, M., Dongwei, J., Pingping, Y., Yongfeng, R.: Energy storage system to improve flexible and stable operation of microgrid in grid-connected and islanding mode. In: 2021 6th International Conference on Power and Renewable Energy (ICPRE), pp. 614–619 (2021).https://doi.org/10.1109/ICPRE52634.2021.9635474
20. Zhong, X., Zhong, W., Liu, Y., Yang, C., Xie, S.: Optimal energy management for multi-energy multi-microgrid networks considering carbon emission limitations. Energy **246**, 123428 (2022). https://doi.org/10.1016/j.energy.2022.123428, https://www.sciencedirect.com/science/article/pii/S0360544222003310
21. Zia, M.F., Benbouzid, M., Elbouchikhi, E., Muyeen, S.M., Techato, K., Guerrero, J.M.: Microgrid transactive energy: Review, architectures, distributed ledger technologies, and market analysis. IEEE Access **8**, 19410–19432 (2020). https://doi.org/10.1109/ACCESS.2020.2968402



(In)approximability of the Maximum 2-Clique Problem for Chordal Graphs

Yuichi Asahiro[1(✉)], Eiji Miyano[2], and Shunta Nakamura[1]

[1] Kyushu Sangyo University, Fukuoka, Japan
asahiro@is.kyusan-u.ac.jp, k25gjk09@st.kyusan-u.ac.jp
[2] Kyushu Institute of Technology, Iizuka, Japan
miyano@ai.kyutech.ac.jp

Abstract. We study the complexity of a distance-based relaxed variant of the maximum clique problem, known as Max 2-Clique: A 2-*clique* in a graph $G = (V, E)$ is a subset $S \subseteq V$ of vertices such that the distance between any pair of vertices in S is at most two in G. Given a graph G with n vertices, the goal of Max 2-Clique is to find a 2-clique of maximum cardinality. It is known that Max 2-Clique cannot be approximated in polynomial time within a factor of $n^{1/2-\varepsilon}$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$. On the positive side, a polynomial-time $O(n^{1/2})$-approximation algorithm exists for Max 2-Clique. In this paper, we establish tight bounds on the approximability of Max 2-Clique for chordal graphs. Specifically, our main contributions are twofold: (i) There exists a polynomial-time $O(n^{1/3})$-approximation algorithm for Max 2-Clique on chordal graphs; and (ii) Max 2-Clique on split graphs cannot be approximated in polynomial time within a factor of $n^{1/3-\varepsilon}$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$, where the class of split graphs is a well-known subclass of the class of chordal graphs.

Keywords: 2-clique · 2-club · Chordal graph · Approximability · Inapproximability

1 Introduction

1.1 Problems

The Maximum Clique Problem (Max Clique) is one of the central problems in graph theory, operations research, and combinatorial optimization. As such, it has attracted significant research attention [5]: Let $G = (V, E)$ be an unweighted undirected graph, where V and E denote the set of vertices and the set of edges, respectively. A *clique* in a graph G is a subset $Q \subseteq V$ of pairwise adjacent vertices. Notably, the decision version of Max Clique was included in Karp's original list of 21 $\mathcal{NP}$-complete problems [14], marking it as a seminal result in

This work was supported by JSPS KAKENHI Grant Numbers JP22K11915 and JP24K02902.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 231–246, 2027.
https://doi.org/10.1007/978-3-032-30521-3_15

computational complexity theory. Cliques have been used to model a wide range of problems across diverse fields, as they represent extremely cohesive groups, i.e., all members are pairwise adjacent. However, this strict requirement can be overly restrictive in many practical applications, where it may suffice for each member to be reachable within at most d hops for some d. To address this, various relaxations of the clique concept have been proposed [4,18,19].

In this paper, we consider a distance-based relaxed variant of Max Clique, known as the maximum distance-d clique problem (Max d-Clique) for a positive integer d [1,15,16]: A *distance-d clique* (*d-clique* for short) in a graph $G = (V, E)$ is a subset $S \subseteq V$ of vertices such that for pairs of vertices $u, v \in S$, the distance between u and v in G is at most d. Given a graph G, the goal of Max d-Clique is to find a d-clique of maximum cardinality in G. Another related problem is the maximum diameter-d club problem (Max d-Club), which resembles but is distinct from Max d-Clique: A *diameter-d club* (*d-club* for short) in a graph G is a subset $S' \subseteq V$ such that a subgraph of G induced by S' has diameter at most d. Given a graph G, the goal of Max d-Club is to find a d-club of maximum cardinality in G. The key difference between Max d-Clique and Max d-Club lies in how distances are measured: In a d-clique, every pair of vertices has distance at most d **in the entire graph** G, whereas in a d-club, the diameter of the induced subgraph is at most d, i.e., every pair of vertices has distance at most d **within the d-club itself**. Thus, a d-club is always a d-clique, but the converse does not necessarily hold, since the shortest path between two vertices included in the d-clique may pass through vertices outside the d-clique.

1.2 Previous Results

In the following we assume that the input graph $G = (V, E)$ has $|V| = n$ vertices and $|E| = m$ edges. For Max Clique, that is, for Max 1-Clique and Max 1-Club, the best known polynomial-time approximation algorithm guarantees a solution within a factor of $n(\log\log n)^2/(\log n)^3$ of the optimum [10]. On the negative side, Max Clique cannot be approximated in polynomial time within a factor of $n^{1-\varepsilon}$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$ [21]. For $d \geq 2$, it is known that both Max d-Clique and Max d-Club cannot be approximated in polynomial time within a factor of $n^{1/2-\varepsilon}$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$ [2]. On the positive side, polynomial-time $O(n^{1/2})$-approximation algorithms have been developed for both Max d-Clique and Max d-Club with $d \geq 2$ [2].

Although Max d-Clique and Max d-Club (including Max Clique) are difficult to approximate as described above, it is known that Max Clique is polynomial-time solvable when the input graph is restricted to certain classes such as planar graphs, chordal graphs, or graphs with bounded degrees [12]. This line of research, i.e., restricting the input to specific graph classes, is a widely studied approach for addressing computationally hard problems. Here we summarize some known results of Max d-Clique and Max d-Club for various graph classes (refer, e.g., to [9] for graph classes and related parameters). Both Max d-Clique and Max d-Club are $\mathcal{NP}$-hard for graphs with degeneracy at

most three when $d \geq 3$, or at most two when $d \geq 5$ [7]. On the positive side, for trees that have representative simple structures, Max d-Club can be solved in polynomial time [11]. Since an optimal solution for Max d-Clique is a d-club for trees, the two problems Max d-Club and Max d-Clique are equivalent on trees. Then, for split graphs, Max 2-Club cannot be approximated in polynomial time within a factor of $O(n^{1/3-\varepsilon})$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$, while a polynomial-time $O(n^{1/3})$-approximation algorithm is known [3]. On the other hand, both Max d-Club and Max d-Clique become trivial for split graphs when $d = 1$ or $d \geq 3$. To the best of our knowledge, whether Max 2-Clique can be solved in polynomial time or not for split graphs is unknown. Since every split graph is a chordal graph, the class of chordal graphs forms a superclass of the class of split graphs. (Note that a tree is also a chordal graph.) Thus, the inapproximability of Max 2-Club on split graphs is transferred to Max 2-Club on chordal graphs; for chordal graphs, Max 2-Club cannot be approximated in polynomial time within a factor of $O(n^{1/3-\varepsilon})$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$. However, whether Max 2-Club on chordal graphs can be approximated within a factor of $O(n^{1/3})$ in polynomial time is also unknown. Besides, Max d-Club is solvable in polynomial time when the input is (i) interval, (ii) strongly chordal, (iii) weakly chordal, or (iv) chordal for odd $d \geq 3$ [13]. It has been also unknown so far whether Max 2-Clique is solvable in polynomial time or not for chordal graphs.

1.3 Contributions

This paper addresses the gaps identified in the previous section for chordal (including split) graphs and $d = 2$. Below, we summarize the obtained results:

- Max 2-Clique and also Max 2-Club can be approximated in polynomial time within a factor of $O(n^{1/3})$ for chordal graphs (and hence also for split graphs).
- Max 2-Clique on split graphs cannot be approximated in polynomial time within a factor of $n^{1/3-\varepsilon}$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$. (Hence the same inapproximability bound also holds for chordal graphs.)

1.4 Organization

Section 2 introduces the notation and problem definitions. In Sect. 3, we present the approximability result for Max 2-Clique on chordal graphs, followed by the corresponding result for Max 2-Club. Section 4 is devoted to establishing the inapproximability of Max 2-Clique on chordal graphs. Finally, Sect. 5 concludes the paper with several open problems.

2 Preliminaries

2.1 Notation

Let $G = (V, E)$ be a connected undirected graph, where V and E are the set of vertices and the set of edges, respectively. $V(G)$ and $E(G)$ also denote the

vertex set and the edge set of G, respectively. We denote an edge with endpoints u and v by $\{u, v\}$. For a vertex v, the set of vertices adjacent to v in G, i.e., the open neighborhood of v is denoted by $N_G(v)$, and $N_G[v]$ denotes the closed neighborhood of v, i.e., $N_G[v] = N_G(v) \cup \{v\}$. We generalize the notions of $N_G(v)$ and $N_G[v]$ to distance two. Let $N_G^2(v)$ denote the set of all vertices that are of distance **exactly** two from a vertex v, that is, $N_G^2(v) = \{w \mid w \in N_G(u) \setminus N_G(v)$ such that $u \in N_G(v)\}$. The set of all vertices that are of distance **at most** two from a vertex v is denoted by $N_G^2[v]$, namely, $N_G^2[v] = N_G[v] \cup N_G^2(v)$. The degree $d_G(v)$ of a vertex v in G is defined as $|N_G(v)|$. Then the maximum degree among all vertices in a graph G is denoted by $\Delta(G)$, i.e., $\Delta(G) = \max_{v \in V(G)}\{d_G(v)\}$.

A graph H is a subgraph of a graph $G = (V, E)$ if $V(H) \subseteq V$ and $E(H) \subseteq E$. For a subset of vertices $U \subseteq V$, $G[U]$ denotes the subgraph of G induced by U, where the edge set of $G[U]$ includes all the edges of G connecting two vertices in U.

A path P of length ℓ from a vertex v_0 to a vertex v_ℓ in G is represented as a sequence of vertices such as $P = \langle v_0, v_1, \ldots, v_\ell \rangle$, where $v_i \in V(G)$ for $0 \leq i \leq \ell$ and $\{v_i, v_{i+1}\} \in E(G)$ for $0 \leq i \leq \ell - 1$. For a pair of vertices u and v in G, the length of a shortest path from u to v, i.e., the distance between u and v is denoted by $dist_G(u, v)$, and the diameter of G is defined as $diam(G) = \max_{u,v \in V}\{dist_G(u, v)\}$.

Let $G = (V, E)$ be a graph and $H = (V', E')$ be its subgraph, where $V' \subseteq V$ and $E' \subseteq E$. Then, the subgraph H or the subset V' of vertices is called a *clique*, if $E' = \{\{u, v\} \mid u, v \in V'\}$. When $E' = \emptyset$, H or V' is called an *independent set*. A graph G is *chordal* if every cycle $\langle v_0, v_1, v_2, \ldots, v_{\ell-1}, v_0 \rangle$ in G of length $\ell \geq 4$ has at least one chord, where a chord is an edge connecting two vertices v_i and v_j such that $j \neq (i + 1) \mod \ell$. If a cycle does not have any chord, it is called an *induced cycle*. A graph G is a *split graph* if there is a partition of $V(G)$ into a clique C and an independent set I such that $C \cap I = \emptyset$ and $C \cup I = V(G)$. Using these C and I, we sometimes write a split graph G as $G = (C, I, E)$, where E is the edge set of G. It is known that every split graph is a chordal graph; the class of split graphs is a subclass of the class of chordal graphs [9].

For the maximization problems, an algorithm `ALG` is called a σ-approximation algorithm and `ALG`'s approximation ratio is σ if $OPT(G)/ALG(G) \leq \sigma$ holds for every input G, where $ALG(G)$ and $OPT(G)$ are the numbers of vertices of obtained subgraphs by `ALG` and an optimal algorithm, respectively. A gap-preserving reduction Γ from a maximization problem Π_1 to another maximization problem Π_2 with four parameters (functions) f_1, f_2, α, and β is defined as follows: Given an instance x of Π_i, Γ computes an instance y of Π_2 in polynomial time such that (i) if $OPT_{\Pi_1}(x) \geq f_1(x)$, then $OPT_{\Pi_2}(y) \geq f_2(y)$; and (ii) if $OPT_{\Pi_1}(x) < f_1(|x|)/\alpha(|x|)$, then $OPT_{\Pi_2}(y) < f_2(y)/\beta(|y|)$, where $OPT_{\Pi_1}(x)$ and $OPT_{\Pi_2}(y)$ are optimal values for the instance x of Π_1 and the instance y of Π_2, respectively. If Π_1 can not be approximated in polynomial time within a factor of $\alpha(|x|)$ unless $\mathcal{P} = \mathcal{NP}$ and a gap-preserving reduction from Π_1 to Π_2

can be constructed, then Π_2 can not be approximated within a factor of $\beta(|y|)$ in polynomial time unless $\mathcal{P} = \mathcal{NP}$ [20].

2.2 Problems

We first provide the formal definition of a d-clique.

Definition 1. *A subgraph H of a graph G is a* d-clique *if $dist_G(u, v) \leq d$ holds for every pair of $u, v \in V(H)$.*

The diameter of a d-clique may be greater than d, and also a d-clique may be disconnected. A d-clique with diameter at most d is called d-club.

Definition 2. *A subgraph H of a graph G is a* d-club *if $dist_H(u, v) \leq d$ holds for every pair of $u, v \in V(H)$, i.e., $diam(H) \leq d$.*

A d-club (or a d-clique) is originally defined as a maximal subgraph in [16], i.e., no super set of vertices of a d-club (or a d-clique) forms a d-club (or a d-clique). However, according to the modified definitions in [6,8], we do not restrict our attention to maximal ones in this paper. Namely, a non-maximal d-club (or a non-maximal d-clique) can also be treated as a d-club (or a d-clique). As for the maximality, it is known to be $\mathcal{NP}$-complete to answer whether a given d-club is maximal or not [17].

The maximum d-clique problem (Max d-Clique), and the maximum d-club problem (Max d-Club) are formulated as follows:

Problem: Maximum d-Clique (Max d-Clique)
Input: A connected undirected graph $G = (V, E)$.
Output: A maximum d-clique in G.

Problem: Maximum d-Club (Max d-Club)
Input: A connected undirected graph $G = (V, E)$.
Output: A maximum d-club in G.

Throughout the paper, we use the following notation: $n = |V|$ and $m = |E|$ for the input graph G. In the case $d = 1$, the above two problems are equivalent; the problems Max 1-Club and Max 1-Clique are identical to Max Clique which is to find a maximum clique in a given graph, and hence they are generalizations of Max Clique in terms of distance and diameter of output subgraphs.

3 Approximability

In this section, we show approximability results. First we show that MAX 2-CLIQUE on chordal graphs can be approximated in polynomial time within a factor of $O(n^{1/3})$ in Sect. 3.1. Then, in Sect. 3.2, we see that MAX 2-CLUB on chordal graphs is also approximated in polynomial time within the same factor $O(n^{1/3})$.

3.1 MAX 2-CLIQUE on Chordal Graphs

This section discusses the approximability of MAX 2-CLIQUE for chordal graphs. Since a maximum clique in a chordal graph can be found in polynomial time [12], we assume that an optimal solution for MAX 2-CLIQUE is not a clique.

Let S be an optimal set of vertices for MAX 2-CLIQUE on chordal graph G such that $G[S]$ does not form a clique. The diameter of $G[S]$ is unclear, since it can be a disconnected graph for MAX 2-CLIQUE. However, for any vertex $r \in S$, $S \subseteq N_G^2[r]$ holds. For simplicity, $G[S]$ is denoted by H. Let $r \in S$ denote a vertex satisfying that there exists a vertex $v \in S$ with $dist_G(r, v) = 2$ (note that it is not "$dist_H(r, v) = 2$").

Let $d = deg_G(r)$, $N_G(r) = \{v_1, v_2, \ldots, v_d\}$, and $N_G^*(r) = N_G(r) \cap S$. Note that $N_G^*(r) = \emptyset$ can happen, since H may be a disconnected 2-clique. Let $N'_G(v_i)$ denote a set of vertices in $N_G^2(r) \cap S$ which belong to $N_G(v_i)$ but not belong to $N_G(v_j) \cap S$ for $1 \leq j \leq i-1$. That is, formally, $N'_G(v_i) = \{w \in N_G(v_i) \cap S \cap N_G^2(r) \mid w \notin \bigcup_{j=1}^{j=i-1} N'_G(v_j)\}$ for $1 \leq i \leq d$. See Fig. 1 for an illustration. We can assume $|N'_G(v_1)| \geq |N'_G(v_2)| \geq \cdots \geq |N'_G(v_d)|$ without loss of generality: To obtain a contradiction, suppose that $|N'_G(v_i)| < |N'_G(v_{i+1})|$, and let G' be the graph obtained by swapping the indices i and $i+1$. Then, in G', $N'_{G'}(v_i) = N'_G(v_{i+1}) \cup T$ and $N'_{G'}(v_{i+1}) = N'_G(v_i) \setminus T$, where T is a set of vertices in $N'_G(v_i) \cap N'_G(v_{i+1})$. Thus, $|N'_{G'}(v_i)| > |N'_{G'}(v_{i+1})|$, since $|N'_G(v_i)| < |N'_G(v_{i+1})|$.

The algorithm we use in this section is `FindStar` which simply finds a vertex u of the maximum degree $\Delta(G)$, and outputs its closed neighborhood $N_G[u]$. The subgraph induced by $N_G[u]$ is clearly a 2-clique (more precisely, a 2-club), and obviously `FindStar` runs in polynomial time. It is known that `FindStar` is an $O(n^{1/2})$-approximation algorithm for MAX 2-CLIQUE (and also for MAX 2-CLUB) on general graphs [2]. We first consider the case that $N'_G(v_2) = \emptyset$.

Lemma 1. *If $N'_G(v_2) = \emptyset$, then the size $|S|$ of the maximum 2-clique is at most $2\Delta(G)$.*

Proof. First, the assumption $N'_G(v_2) = \emptyset$ implies that $N'_G(v_i) = \emptyset$ also for $i \geq 3$ by the assumption $|N'_G(v_i)| \geq |N'_G(v_{i+1})|$.

Since $|N'_G(v_1)| \leq \Delta(G) - 1$,

$$\begin{aligned} |S| &\leq |\{r\} \cup N_G(r) \cup N'_G(v_1)| \\ &\leq |\{r\} \cup N_G(r) \cup (N_G(v_1) \setminus \{r\})| \\ &\leq 1 + \Delta(G) + (\Delta(G) - 1) \\ &= 2\Delta(G), \end{aligned}$$

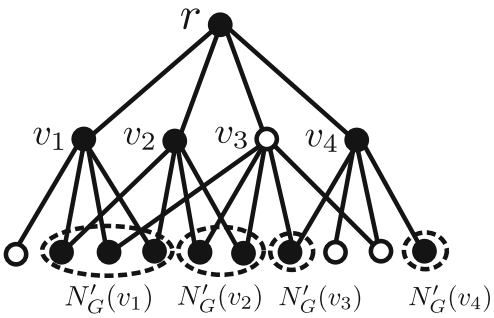


Fig. 1. A solution S is represented by filled vertices. Within S, there exists a vertex whose distance from a designated vertex r is two. This figure highlights the vertex r along with $N^2_G[r]$, where $N_G(r) = \{v_1, v_2, v_3, v_4\}$ and $N^*_G(r) = \{v_1, v_2, v_4\}$. Dashed circles indicate the sets $N'_G(v_1)$ through $N'_G(v_4)$. Note that only the edges between r and $N_G(r)$, and between $N_G(r)$ and $N^2_G(r)$, are shown. As a result, the subgraph induced by the filled vertices may not appear to form a 2-clique in the figure.

where the first inequality follows from the facts that $N_G(r) \cap \mathrm{S} \subseteq N_G(r)$ and $N'_G(v_i) = \emptyset$ for $i \geq 2$. □

Next, we consider the case $N'_G(v_2) \neq \emptyset$ in the following, i.e., there are at least two v_i's such that $|N'_G(v_i)| > 0$. The following two lemmas demonstrate existence of certain types of edges.

Lemma 2. *Suppose that $u_i \in N^2_G(r) \cap N_G(v_i)$ and $u_j \in N^2_G(r) \cap N_G(v_j)$. If $\{u_i, u_j\} \in E(G)$, then $E(G)$ includes either or both $\{v_i, u_j\}$ and $\{v_j, u_i\}$.*

Proof. If $v_i = v_j$, then $\{v_j, u_j\} = \{v_i, u_j\}$ and $\{v_i, u_i\} = \{v_j, u_i\}$. By definition $u_j \in N^2_G(r) \cap N_G(v_j)$, $\{v_j, u_j\} \in E(G)$. Analogously, $\{v_i, u_i\}$ exists in $E(G)$. Thus, the lemma holds.

Consider the case $v_i \neq v_j$. For obtaining a contradiction, assume that neither $\{v_i, u_j\}$ nor $\{v_j, u_i\}$ exists in a chordal graph G. There is a cycle $C = \langle r, v_i, u_i, u_j, v_j, r\rangle$ of length five in G, and neither $\{r, u_i\}$ nor $\{r, u_j\}$ exists in G by definitions of u_i and u_j. See Fig. 2 for an illustration. Thus, we can have $\{v_i, v_j\}$ only as a chord of the cycle C. Even though, still there is an induced cycle $\langle v_i, u_i, u_j, v_j, v_i\rangle$ of length four. Namely, by the assumption, there is an induced cycle of length four or five, implying that G is not chordal. This is a contradiction, and thus $E(G)$ includes either or both $\{v_i, u_j\}$ and $\{v_j, u_i\}$. □

Lemma 3. *Suppose that $u_i \in N^2_G(r) \cap N_G(v_i)$ and $u_j \in N^2_G(r) \cap N_G(v_j)$. If there is a path $\langle u_i, w, u_j\rangle$, for some $w \notin N_G[r]$ and $\{u_i, u_j\} \notin E(G)$, then $E(G)$ includes both $\{w, v_i\}$ and $\{w, v_j\}$.*

Proof. If $v_i = v_j$, then there is a cycle $C = \langle v_i, u_i, w, u_j, v_i\rangle$. Since $\{u_i, u_j\} \notin E(G)$ and G is chordal, $\{w, v_i\}(= \{w, v_j\})$ must exist as a chord of C.

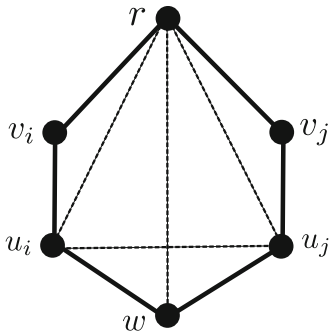


Fig. 2. Cycle $\langle r, v_i, u_i, u_j, v_j, r\rangle$ of length five. Dashed lines indicate where edges can not exist by assumption of the proof of Lemma 2.

Consider the case $v_i \neq v_j$. First, there exists a cycle $C = \langle r, v_i, u_i, w, u_j, v_j, r\rangle$ in G, where none of $\{r, u_i\}$, $\{r, u_j\}$, $\{r, w\}$, and $\{u_i, u_j\}$ exists. See Fig. 3 for an illustration. Look at the subpath $\langle v_i, u_i, w, u_j\rangle$ of C, whose length is three. To obtain a contradiction, assume that $\{w, v_i\} \notin E(G)$. Even if we add all remaining (possible) edges $\{v_i, v_j\}$, $\{v_i, u_j\}$, $\{u_i, v_j\}$, and $\{w, v_j\}$, there exists an induced cycle $\langle v_i, u_i, w, u_j, v_i\rangle$ of length four. This contradicts the assumption that G is chordal. Hence $\{w, v_i\} \in E(G)$. By a similar argument, $\{w, v_j\} \in E(G)$, too. □

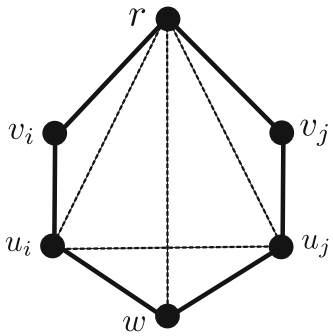


Fig. 3. Cycle $\langle r, v_i, u_i, w, u_j, v_j, r\rangle$ of length six. Dashed lines indicate where edges can not exist by assumption of the proof of Lemma 3.

Next, we bound the size of $\bigcup_{j=1}^{d} N'_G(v_j)$, which plays an important role to show the approximability of `FindStar`. Let $q = |\bigcup_{j=1}^{d} N'_G(v_j)|$.

Lemma 4. *If* $N'_G(v_2) \neq \emptyset$, *then* $q \leq 4 \cdot \Delta(G)^{3/2} + 1$.

Proof. For simplicity, let $V = V(G)$, $E = E(G)$, and $\Delta = \Delta(G)$. We estimate q by analyzing the number of edges between $N_G(r)$ and $N_G^2(r)$.

The proof strategy. The proof proceeds as follows.

(I) For a pair of vertices $u_i \in N'_G(v_i)$ and $u_j \in N'_G(v_j)$, we *assign* the pair (u_i, u_j) to an existing edge between $N_G(r)$ and $N^2_G(r)$, utilizing Lemmas 2 and 3. See Fig. 4 for an illustration.

(II) A single edge may be assigned to multiple such pairs by (I). We observe that the number of pairs assigned to any edge is at most 8Δ.

(III) Since the number of edges emanating from $N_G(r)$ is at most $\Delta \cdot |N_G(r)|$, the total number $q(q-1)/2$ of pairs in $\bigcup_{j=1}^{d} N'_G(v_j)$ is bounded above by $8\Delta^2 \cdot |N_G(r)|$. This yields the desired bound on q.

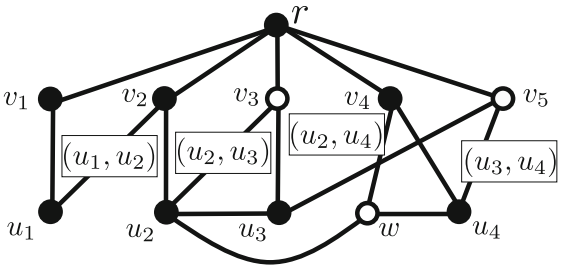


Fig. 4. Assignment of each pair of vertices in $N^2_G(r) \cap S$ to an edge between $N_G(r)$ and $N^2_G(r)$ in the proof of Lemma 4. The pairs (u_1, u_2), (u_2, u_3), (u_2, u_3), and (u_3, u_4) are assigned to the edges $\{v_2, u_1\}$, $\{v_3, u_2\}$, $\{v_4, w\}$, and $\{v_5, u_4\}$, respectively. Unfilled vertices are assumed not to be included in the solution. However, this assumption is immaterial, since the proof relies only on the edges between $N_G(r)$ and $N^2_G(r)$.

(I) We begin by selecting an edge to which the pair (u_i, v_j) will be assigned, where $i < j$. To ensure that $dist_G(u_i, u_j) \leq 2$, there must exist additional edges beyond the four fixed ones: $\{r, v_i\}$, $\{v_i, u_i\}$, $\{r, v_j\}$, and $\{v_j, u_j\}$. Note that neither $\{r, u_i\}$ nor $\{r, u_j\}$ can exist; otherwise $u_i \in N_G(r)$ or $u_j \in N_G(r)$, contradicting the assumptions that $u_i \in N'_G(v_i)$ and $u_j \in N'_G(v_j)$. Therefore, to satisfy the distance condition, one of the following must hold: (A) $\{v_i, u_j\} \in E$, (B) $\{v_j, u_i\} \in E$, (C) $\{u_i, u_j\} \in E$, or (D) there exists a path $\langle u_i, w, u_j \rangle$ for some vertex $w \in V \setminus \{r, v_i, v_j\}$. (Note that, for example, if $w = v_j$ then $\{u_i, v_j\} \in E$). In the following, we examine each of these cases (A)-(D) to determine an appropriate edge to which the pair (u_i, u_j) can be assigned.

(A) If $\{v_i, u_j\} \in E$, it contradicts the assumption $u_j \in N'_G(v_j)$ for $i < j$ by the definition of $N'_G(v_j)$. Thus, we assume $\{v_i, u_j\} \notin E$ in the following.

(B) If ($\{v_i, u_j\} \notin E$ and) $\{v_j, u_i\} \in E$, we assign the pair (u_i, u_j) to this edge $\{v_j, u_i\}$. By this type, pairs (u_i, u) for $u \in N'_G(v_j)$ may be assigned to $\{v_j, u_i\}$. (In Fig. 4, the pair (u_1, u_2) is assigned to the edge $\{v_2, u_1\}$.)

(C) Suppose that ($\{v_i, u_j\} \notin E$ and) $\{u_i, u_j\} \in E$. From Lemma 2, $\{\{v_i, u_j\}, \{v_j, u_i\}\} \cap E \neq \emptyset$. However, from (A), $\{v_i, u_j\} \notin E$, and thus $\{v_j, u_i\} \in E$. Therefore, this case is included in (B), and hence we assign the pair (u_i, u_j) to this edge $\{v_j, u_i\}$. (In Fig. 4, the pair (u_2, u_3) is assigned to the edge $\{v_3, u_2\}$.)

(D) Suppose that ($\{v_i, u_j\} \notin E$, $\{v_j, u_i\} \notin E$, $\{u_i, u_j\} \notin E$, and) a path $\langle u_i, w, u_j \rangle$ for some vertex $w \in V \setminus \{r, v_i, v_j\}$ exists. Note that $w \notin S$ can happen. We further distinguish this case into five subcases, (D-1) $w = v_k$ for some $k \notin \{i, j\}$, (D-2) $w \in N_G^2(r) \cap N_G(v_k)$ for some $k \notin \{i, j\}$, (D-3) $w \in N_G^2(r) \cap N_G(v_i)$, (D-4) $w \in N_G^2(r) \cap N_G(v_j)$, and (D-5) $w \notin N_G^2[r]$.

(D-1) The case $w = v_k$ for some $k \notin \{i, j\}$: See Fig. 5 for an illustration. We assign the pair (u_i, u_j) to the second edge $\{v_k, u_j\}$ of the path $\langle u_i, v_k, u_j \rangle$. (In Fig. 4, the pair (u_3, u_4) is assigned to the edge $\{v_5, u_4\}$.)

(D-2) The case $w \in N_G^2(r) \cap N_G(v_k)$ for some $k \notin \{i, j\}$: See Fig. 6 for an illustration. Look at the second edge $\{w, u_j\}$ of the path $\langle u_i, w, u_j \rangle$. From Lemma 2, there must exist $\{v_k, u_j\}$ and/or $\{v_j, w\}$ in E. If $\{v_k, u_j\} \in E$, we assign the pair (u_i, u_j) to $\{v_k, u_j\}$, otherwise, we assign the pair (u_i, u_j) to $\{w, v_j\}$.

(D-3) The case $w \in N_G^2(r) \cap N_G(v_i)$: See Fig. 7 for an illustration. As in (D-2), from Lemma 2, there must exist $\{v_j, w\}$ in E.

(D-4) The case $w \in N_G^2(r) \cap N_G(v_j)$: In this case, the edge $\{v_j, w\}$ exists in E. Thus, we assign the pair (u_i, u_j) to the edge $\{v_j, w\}$.

(D-5) The case $w \notin N_G^2[r]$: See Fig. 3 again for an illustration. By Lemma 3, there exists an edge $\{v_j, w\}$ in E, i.e., $w \in N_G(v_j)$. This contradicts the assumption $w \notin N_G^2[r]$, that is, this case does not occur.

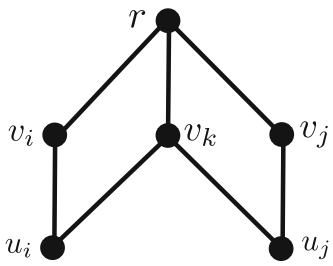


Fig. 5. (D-1) A path $\langle u_i, w, u_j \rangle$, where $w = v_k$ for some $k \notin \{i, j\}$.

By the above cases (A)-(D), every pair (u_i, u_j) for $u_i \in N'_G(v_i)$ and $u_j \in N'_G(v_j)$ for $i < j$ is assigned to an edge that emanates from $N_G(r)$.

(II) Next we estimate how many pairs can be assigned to a single edge between $N_G(r)$ and $N_G^2(r)$. Consider an edge $\{v_i, u_i\}$, where $v_i \in N_G(r)$ and $u_i \in N'_G(v_i)$. Recall that Case (A) does not assign any pair to $\{v_i, u_i\}$, and Case (C) is subsumed by Case (B). Therefore, we focus on Cases (B) and (D) in the following analysis.

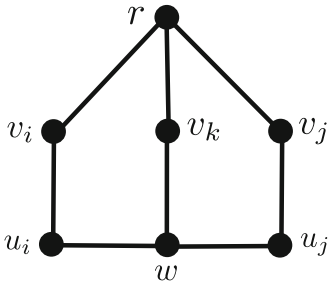


Fig. 6. (D-2) A path $\langle u_i, w, u_j \rangle$, where $w \in N_G^2(r) \cap N_G(v_k)$ for some $k \notin \{i, j\}$. From Lemma 2, $\{v_k, u_j\} \in E$ and/or $\{v_j, w\} \in E$.

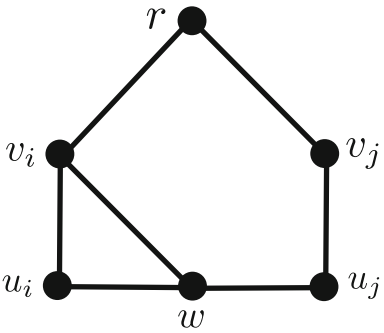


Fig. 7. (D-3) A path $\langle u_i, w, u_j \rangle$, where $w \in N_G^2(r) \cap N_G(v_i)$. From Lemma 2, $\{v_j, w\} \in E$.

(B) When $\{v_j, u_i\} \in E$, the pair (u_i, u_j) is assigned to the edge $\{v_j, u_i\}$: In this case, multiple pairs (u_i, u_h), where $h \in N'_G(v_j)$, may be assigned to the same edge $\{v_j, u_i\}$. Since $deg_G(v_j) \leq \Delta$, the number of pairs assigned to $\{v_j, u_i\}$ under Case (B) is at most Δ.

(D) As in the above discussion, we consider each of the distinguished cases:

(D-1) When $\{v_k, u_j\}$ is the second edge of a path $\langle u_i, v_k, u_j \rangle$, the pair (u_i, u_j) is assigned to the edge $\{v_k, u_j\}$: This edge $\{v_k, u_j\}$ may serve as the second edge of such a path $\langle u, v_k, u_j \rangle$ for some $u \in N_G^2(r) \cap S$. Since $deg_G(v_k) \leq \Delta$, the number of pairs assigned to $\{v_k, u_j\}$ under Case (D-1) is at most Δ.

(D-2) When $\{w, u_j\}$ is the second edge of a path $\langle u_i, w, u_j \rangle$, where $w \in N'_G(v_k)$ and $k \notin \{i, j\}$, the pair (u_i, u_j) is assigned to either $\{v_k, u_j\}$ or $\{v_j, w\}$: The edge $\{w, u_j\}$ may serve as the second edge of a path $\langle u, w, u_j \rangle$ or of a path $\langle u, u_j, w \rangle$ when $w \in S$, for some $u \in N_G^2(r) \cap S$. Since $deg_G(w) \leq \Delta$ and $deg_G(u_j) \leq \Delta$, at most 2Δ pairs can be assigned to each of the edges $\{v_k, u_j\}$ and $\{w, v_j\}$ under Case (D-2).

(D-3) When $\{w, u_j\}$, with $w \in N'_G(v_i)$, is the second edge of a path $\langle u_i, w, u_j \rangle$, the pair (u_i, u_j) is assigned to the edge $\{v_j, w\}$: Similar to

Case (D-2), the edge $\{w, u_j\}$ may serve as the second edge of at most 2Δ paths of the form $\langle u, w, u_j\rangle$ and $\langle u, u_j, w\rangle$ (when $w \in S$), for some $u \in N_G^2(r) \cap S$. Thus, at most 2Δ pairs are assigned to the edge $\{v_j, w\}$ under Case (D-3).

(D-4) When $\{w, u_j\}$, with $w \in N'_G(v_j)$, is the second edge of a path $\langle u_i, w, u_j\rangle$, the pair (u_i, u_j) is assigned to the edge $\{v_j, w\}$: Similar to Cases (D-2) and (D-3), at most 2Δ pairs are assigned to the edge $\{v_j, w\}$ under Case (D-4).

In total, the number of pairs assigned to any single edge between $N_G(r)$ and $N_G^2(r)$ is at most

$$\Delta + \Delta + 2\Delta + 2\Delta + 2\Delta = 8\Delta,$$

where the first two terms Δ come from Cases (B) and (D-1), and the remaining three terms 2Δ correspond to Cases (D-2), (D-3), and (D-4), respectively.

(III) Since $|N_G(r)| \leq \Delta$, the number of edges between $N_G(r)$ and $N_G^2(r)$ is at most $|N_G(r)| \cdot \Delta \leq \Delta^2$. The number of pairs (u, u')'s, where $u, u' \in \bigcup_{j=1}^{d} N'_G(v_j)$, is $q(q-1)/2$. Thus, it must hold (as a necessary condition)

$$\frac{(q-1)^2}{2} \leq \frac{q(q-1)}{2} \leq \Delta^2 \cdot 8\Delta,$$

which implies

$$q \leq 4 \cdot \Delta^{3/2} + 1.$$

□

Based on the above lemma, we obtain an upper bound of the size of the maximum 2-clique:

Corollary 1. *The size of the maximum 2-clique in a chordal graph G is at most $4 \cdot \Delta(G)^{3/2} + \Delta(G) + 2$.*

Proof. Let G be an input chordal graph, and S be the vertex set of the maximum 2-clique in G. If $N'_G(v_2) = \emptyset$, then $|S| \leq 2\Delta(G)$ from Lemma 1. If $N'_G(v_2) \neq \emptyset$, then

$$\begin{aligned}
|S| &= \left| \{r\} \cup N_G^*(r) \cup \bigcup_{j=1}^{d} N'_G(v_j) \right| \\
&\leq \left| \{r\} \cup N_G(r) \cup \bigcup_{j=1}^{d} N'_G(v_j) \right| \\
&\leq 1 + \Delta(G) + q \\
&\leq 4 \cdot \Delta(G)^{3/2} + \Delta(G) + 2,
\end{aligned}$$

where the first inequality is based on the fact $N_G^*(r) \subseteq N_G(r)$, the second inequality holds since $|\{r\} \cup N_G(r)| \leq 1 + \Delta(G)$, and then the last inequality is from Lemma 4. The lemma follows from the observation that $2\Delta(G) \leq 4 \cdot \Delta(G)^{3/2} + \Delta(G) + 2$. □

Based on the above lemma, we can show that `FindStar` is an $O(n^{1/3})$-approximation algorithm:

Theorem 1. *`FindStar` is an $O(n^{1/3})$-approximation algorithm for* Max 2-Clique *if the input is a chordal graph.*

Proof. Let S^* denote an optimal set of vertices, and let $\Delta = \Delta(G)$. Then, `FindStar` outputs a 2-clique of $\Delta + 1$ vertices.

Consider the case $\Delta \geq n^{2/3}$. Since it is obvious that $|S^*| \leq n$, the approximation ratio of `FindStar` is at most $n/(\Delta + 1) \leq n^{1/3}$.

Next, consider the case $\Delta < n^{2/3}$. It holds $|S^*| \leq 4 \cdot \Delta^{3/2} + \Delta + 2$ from Corollary 1, and thus, the approximation ratio for this case is at most

$$\frac{4 \cdot \Delta^{3/2} + \Delta + 2}{\Delta + 1} = O(\Delta^{1/2}) = O(n^{1/3}).$$

Therefore, in every case, the approximation ratio of `FindStar` is $O(n^{1/3})$. □

3.2 Max 2-Club on chordal graphs

Since a 2-club is a 2-clique, the maximum 2-clique is larger than the maximum 2-club. Thus, the upper bound of the maximum 2-clique size shown in Corollary 1 is transferred to Max 2-Club:

Corollary 2. *The size of the maximum 2-club in a chordal graph G is at most $4 \cdot \Delta(G)^{3/2} + \Delta(G) + 2$.*

Then, the proof of Theorem 1 also gives the next theorem:

Theorem 2. *`FindStar` is an $O(n^{1/3})$-approximation algorithm for* Max 2-Club *if the input is a chordal graph.*

4 Inapproximability

This section investigates the intractability of Max 2-Clique on chordal graphs. It is known that Max 2-Club on split graphs cannot be approximated in polynomial time within a factor of $O(n^{1/3-\varepsilon})$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$ [3]. Since every split graph is a chordal graph, the inapproximability result directly extends to chordal graphs: Max 2-Club on chordal graphs also cannot be approximated in polynomial time within a factor of $O(n^{1/3-\varepsilon})$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$. In the previous section, we showed that both Max 2-Club and Max 2-Clique admit polynomial-time $O(n^{1/3})$-approximation algorithms for chordal graphs. Thus, the remaining task in this section is to show that Max 2-Clique on chordal graphs cannot be approximated in polynomial time within a factor of $O(n^{1/3-\varepsilon})$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$. To this end, we prove the same inapproximability bound, $O(n^{1/3-\varepsilon})$ for split graphs, which immediately implies the corresponding inapproximability result for chordal graphs as well.

The next lemma shows that a maximum 2-clique in a split graph $G = (C, I, E)$ contains C as its vertex subset.

Lemma 5. *A maximum 2-clique in a split graph $G = (C, I, E)$ contains C.*

Proof. Let $S^* = C^* \cup I^*$ be a maximum 2-clique in a split graph $G = (C, I, E)$ such that $C^* \subseteq C$ and $I^* \subseteq I$. For the sake of contradiction, assume $C - C^* \neq \emptyset$. If $I^* = \emptyset$, it contradicts the assumption that S^* is a maximum 2-clique, since C is a clique and thus it is a 2-clique including more vertices than S^*. Hence, if $I^* = \emptyset$, $C^* = C$. Assume that $I^* \neq \emptyset$ in the following.

Let $v \in C \setminus C^*$. Since C is a clique, we have $dist_G(v, u) = 1$ for any $u \in C^*$. Now consider a vertex $u \in I^*$. If $v \in N_G(u)$, then clearly $dist_G(v, u) = 1$. Assume instead that $v \notin N_G(u)$, and let $v' \in C$ be a vertex in $N_G(u)$. Because C is a clique, we have $\{v', v\} \in E(G)$, and hence $dist_G(v, u) = 2$. Therefore, $S^* \cup \{v\}$ is also a 2-clique with $(|S^*| + 1)$ vertices, contradicting the assumption that S^* is a maximum 2-clique. Consequently, the assumption $C - C^* \neq \emptyset$ is false, and thus $C^* = C$ holds. □

The next lemma shows that a maximum 2-clique is a 2-club in a split graph.

Lemma 6. *A maximum 2-clique in a split graph is a 2-club.*

Proof. Let $S^* = C \cup I^*$ be a maximum 2-clique in a split graph $G = (C, I, E)$ such that $I^* \subseteq I$, where S^* contains C, based on Lemma 5. Also let $H = G[S^*]$.

If $|I^*| = 0$, i.e., $I^* = \emptyset$, then $H = G[C]$ is a clique and hence a 2-club. Suppose that $|I^*| = 1$, and let $I^* = \{u\}$. Consider any vertex $v \in C$. If $v \in N_G(u)$, then $dist_H(v, u) = 1$. Otherwise if $v \notin N_G(u)$, there is a vertex $v' \in N_G(u)$ $(\subseteq C)$ such that there is a path $\langle u, v', v\rangle$ of length two in H, i.e., $dist_H(u, v) = 2$. Also, since C is a clique, for any $u, v \in C$, $dist_H(u, v) = 1$. Thus, the distance between any pair of two vertices is at most two in H, i.e., H is a 2-club.

As the last case, assume that $|I^*| \geq 2$. It holds $dist_H(u, v) \leq 2$, if either or both u and v are in C, according to the discussion for the case $|I^*| = 1$. Now, we consider two distinct vertices $u, v \in I^*$. By the definition of $I^* \subseteq I$, there is no edge between u and v. Since H is a 2-clique, there is a vertex w in G such that there is a path $\langle u, w, v\rangle$ of length two in G. It is clear that $w \in C$. Since $C \subseteq S^*$, $w \in S^*$ which ensures that the path $\langle u, w, v\rangle$ also exists in H implying that $dist_H(u, v) = 2$. Therefore H is a 2-club. □

Based on the above lemma, we show the following theorem.

Theorem 3. *In a split graph, a maximum 2-club is also a maximum 2-clique, and vice versa.*

Proof. Let H and H' be a maximum 2-club and a maximum 2-clique in a split graph, respectively. From Lemma 6, H' is also a 2-club. Thus, $|H'| \leq |H|$ holds. It is clear that a 2-club is also a 2-clique, which implies that $|H| \leq |H'|$. Therefore $|H| = |H'|$ holds, i.e., H' (or H) is a maximum 2-club (or a maximum 2-clique), too. □

Since finding a maximum 2-clique is equivalent to finding a maximum 2-club in a split graph from the above theorem, the inapproximability of Max 2-Club on split graphs [3] is transferred to Max 2-Clique on split graphs, which also gives an inapproximability of Max 2-Clique on chordal graphs.

Corollary 3. Max 2-Clique *on split graphs (a subclass of chordal graphs) cannot be approximated in polynomial time to within a factor of* $n^{1/3-\varepsilon}$ *for any* $\varepsilon > 0$ *unless* $\mathcal{P} = \mathcal{NP}$.

5 Conclusion

We have shown that Max 2-Clique on chordal graphs admits a polynomial-time $O(n^{1/3})$-approximation algorithm. On the negative side, we also proved that Max 2-Clique on chordal graphs cannot be approximated in polynomial time within a factor of $n^{1/3-\varepsilon}$ for any $\varepsilon > 0$ unless $\mathcal{P} = \mathcal{NP}$. As a secondary result, we further established that Max 2-Club also admits a polynomial-time $O(n^{1/3})$-approximation algorithm for chordal graphs.

We conclude by listing several open problems:

- Is Max d-Clique on chordal graphs $\mathcal{NP}$-hard for $d \geq 3$?
  - It is known that Max d-Club on chordal graphs can be solved in polynomial time for odd $d \geq 3$ [13], and is $\mathcal{NP}$-hard for even $d \geq 4$ [3].
- What is the complexity of Max 2-Clique on bipartite graphs?
  - While Max 2-Club can be solved in polynomial time for bipartite graphs [3], it remains open whether Max 2-Clique is also polynomial-time solvable.
- Are there graph classes for which the complexities of Max d-Clique and Max d-Club differ?
  - To the best of our knowledge, the complexities of these two problems coincide for all graph classes studied so far (from the viewpoint of the polynomial-time solvability). Bipartite graphs may serve as a candidate for identifying such a separation.

References

1. Alba, R.: A graph-theoretic definition of a sociometric clique. J. Math. Sociol. **3**, 113–126 (1973)
2. Asahiro, Y., Doi, Y., Miyano, E., Samizo, K., Shimizu, H.: Optimal approximation algorithms for maximum distance-bounded subgraph problems. Algorithmica **80**(6), 1834–1856 (2018)
3. Asahiro, Y., Miyano, E., Samizo., K.: Approximating maximum diameter-bounded subgraphs. In: LATIN 2010, LNCS, vol. 6034, pp.616–627(2010)
4. Abello, J., Resende, M.G.C., Sudarsky, S.: Massive Quasi-Clique detection. In: Rajsbaum, S. (ed.) LATIN 2002. LNCS, vol. 2286, pp. 598–612. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45995-2_51
5. Bomze, I.M., Budinich, M., Pardalos, P.M., Pelillo, M.: The maximum clique problem. In: Handbook of Combinatorial Optimization, pp.1–74, Kluwer Academic(1999)
6. Balasundaram, B., Butenko, S., Trukhanov, S.: Novel approaches for analyzing biological networks. J. Combinat. Optim. **10**(1), 23–39 (2005)

7. Baril, A., Castillon, A., Oijid, N.: On the parameterized complexity of non-hereditary relaxations of clique. Theor. Comput. Sci. **1003**, Article 114625(2024)
8. Bourjolly, J.-M., Laporte, G., Pesant, G.: An exact algorithm for the maximum k-club problem in an undirected graph. Europ. J. Oper. Res. **138**, 21–28 (2002)
9. Brandstädt, A. Le, V.B., Spinrad, J.P.: Graph Classes: A Survey, SIAM (1999)
10. Feige, U.: Approximating maximum clique by removing subgraphs. SIAM J. Discr. Math. **18**, 219–225 (2004)
11. Feránndez-Zepeda, J.A., Flores-Lamas, A., Hague, M., Trejo-ánchez, J.A.: A dynamic programming algorithm for the maximum s-club problem on trees. Europ. J. Oper. Res. **329**(2), 426–435 (2026)
12. Gavril, F.: Algorithms for minimum coloring, maximum clique, minimum covering by cliques, and maximum independent set of a chordal graph. SIAM J. Comput. **1**(2), 180–187 (1972)
13. Golovach, P.A., Heggernes, P., Kratsch, D., Rafiey, A.: Finding clubs in graph classes. Discret. Appl. Math. **174**, 57–65 (2014)
14. Karp, R.M.: Reducibility among combinatorial problems. Complex. Comput. Comput. 85–103(1972)
15. Luce, R.D.: Connectivity and generalized cliques in sociometric group structure. Psychometrika **15**(2), 169–190 (1950)
16. Mokken, R.J.: Cliques, clubs and clans. Qual. Quant. **13**, 161–173 (1979)
17. Pajouh, F.M., Balasundaram, B.: On inclusionwise maximal and maximum cardinality k-clubs in graphs. Discret. Optim. **9**, 84–97 (2012)
18. Seidman, S.B.: Network structure and minimum degree. Social Netw. **5**(3), 269–287 (1983)
19. Seidman, S.B., Foster, B.L.: A graph-theoretic generalization of the clique concept. J. Mathematical Sociology **6**(1), 139–154 (1978)
20. Vazirani, V.V.: Approximation Algorithms. Springer (2003)
21. Zuckerman, D.: Linear degree extractors and the inapproximability of max clique and chromatic number. Theory of Computing **3**, 103–128 (2007)

Evaluating AI-Generated Surrogate Optimization Algorithms for Expensive Black-Box Optimization

Rommel G. Regis(✉)

Saint Joseph's University, Philadelphia, PA 19131, USA
rregis@sju.edu

Abstract. Generative AI tools have become adept at writing code for various software development tasks. This paper evaluates the quality of the algorithms and Matlab codes for surrogate optimization that were generated by ChatGPT 5.2. The AI was prompted to write Matlab codes that are suitable for computationally expensive black-box optimization, particularly Bayesian optimization and radial basis function (RBF) methods. The AI-generated Matlab codes for surrogate optimization were compared with an original human-generated Matlab implementation of an established RBF method called Multistart Local Metric Stochastic RBF (MLMSRBF) on eight widely used test problems with up to 32 dimensions and on a 36-dimensional environmental engineering application. The AI-generated Matlab codes are well-structured and they implement and combine well-known surrogate optimization approaches. Among the codes it generated was a somewhat novel RBF algorithm that uses a strategy commonly associated with Bayesian optimization. Moreover, the performances of the AI-generated TuRBO-style Bayesian optimization code and the AI-generated Stochastic RBF-style code are comparable with the performance of the MLMSRBF code on the test problems. This suggests that AI tools can produce high quality surrogate optimization codes and have great potential to assist in extending or developing novel algorithms for expensive black-box optimization.

Keywords: Surrogate optimization · black-box optimization · generative AI · Bayesian optimization · radial basis functions

1 Introduction

Generative AI tools are increasingly being used to solve problems in Operations Research and Mathematical Optimization [23]. In particular, ChatGPT has been used to formulate problems in engineering optimization [21]. It has been used to develop solution approaches for variants of the Shortest Path Problem and to write Python codes that implement the algorithms [10]. ChatGPT has also been used to generate a C code that implements a swarm intelligence algorithm to

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 247–265, 2027.
https://doi.org/10.1007/978-3-032-30521-3_16

solve univariate global optimization problems [17]. This paper considers ChatGPT for generating codes that implement algorithms for expensive black-box optimization.

Due to the ubiquity of expensive simulation-based black-box optimization problems in many engineering applications, there is great demand for surrogate optimization algorithms. These methods use surrogate models that approximate the objective and/or constraint functions in every iteration and uses these surrogates to guide the selection of subsequent points for function evaluations. Popular methods include Bayesian optimization [2,6,7,9], which uses Gaussian Process (GP) surrogates [13], and radial basis function methods [8,15]. While there are some commercial and publicly available codes that implement surrogate methods, they are sometimes only available in one or two programming languages. For example, the TuRBO algorithm [5] that implements local Bayesian optimization is available in Python. The SACOBRA algorithm [3], which uses radial basis function (RBF) interpolation for constrained black-box optimization, is available in R. The Stochastic RBF algorithm was written in Matlab and its variants are also available in Python through the Surrogate Optimization Toolbox (pySOT) [4]. Moreover, commercial packages such as BayesOpt and surrogateopt require Matlab's Statistics and Machine Learning Toolbox [20] and the Global Optimization Toolbox [19], respectively. Potential users who need one of these algorithms might find that it is not available in their preferred language and could resort to AI tools to code the algorithm. Additionally, other users who need to solve expensive black-box optimization problems but are not be familiar with the literature might prompt an AI tool to generate such a method. As generative AI tools become more adept at writing codes, it would be interesting to explore how well AI tools can generate codes that implement algorithms for expensive black-box optimization, particularly surrogate optimization methods.

This paper explores the capability of ChatGPT 5.2 to generate Matlab codes for solving expensive black-box optimization problems. When prompted for such a code, GPT-5.2 generated one that implements Bayesian optimization with the classic Expected Improvement criterion [9]. It was also prompted to generate alternative methods that use radial basis function (RBF) surrogates and a local variant of Bayesian optimization that is more suitable for higher dimensional problems. One of the AI-generated codes was a hybrid of ideas from Bayesian optimization and RBF methods producing a somewhat novel surrogate optimization algorithm. The AI-generated Matlab codes for surrogate optimization were compared with an original human-generated Matlab implementation of an established RBF method called Multistart Local Metric Stochastic RBF (MLMSRBF) [15] on widely used test problems. The Stochastic RBF method was chosen for comparison since the paper that introduced it was recently given the INFORMS Journal on Computing Test of Time Award for 2005–2009 [16]. The Matlab codes generated by GPT-5.2 are well-structured and well-documented and implement well-known surrogate optimization approaches. The AI-generated Stochastic RBF-style code and the AI-generated TuRBO-style Bayesian optimization code yielded comparable results to those obtained by the non-AI MLMSRBF

code on several test problems, including an environmental engineering application. This suggests that AI tools such as GPT-5.2 can produce codes for surrogate optimization with comparable performance to a human-generated code.

2 Surrogate Optimization Methods

In this study, we evaluate the AI-generated Matlab codes for bound-constrained expensive black-box optimization problems, which have the form:

$$\begin{array}{c} \min \ f(x) \\ \text{s.t.} \\ \ell \leq x \leq u \\ x \in \mathbb{R}^d \end{array} \tag{1}$$

where $\ell, u \in \mathbb{R}^d$ and the objective functions f is black-box and computationally expensive. A similar experiment can be done for optimization problems with constraints, but this will be the subject of future work.

Surrogate optimization methods for problem (1) typically follow the same structure (Algorithm 1). Below, n_0 is the number of initial sample points, n_{max} is the maximum number of function evaluations, $\mathcal{P}_n = \langle x_1, x_2, \ldots, x_n \rangle$ denotes the ordered set of the first n sample points, and $f(\mathcal{P}_n) = \langle f(x_1), f(x_2), \ldots, f(x_n) \rangle$ is the corresponding ordered set of function values.

Algorithm 1 Surrogate Optimization Framework

Inputs: $n_0 \geq d+1$, $n_{max} > n_0$
Choose n_0 sample points $\mathcal{P}_{n_0} = \langle x_1, \ldots, x_{n_0} \rangle \subset [\ell, u]$ ▷ space-filling design
Evaluate f at the points in $\mathcal{P}_{n_0}$ and set $n \leftarrow n_0$
while $n < n_{max}$ **do**
  Fit surrogate model $s_n(x)$ using data from $\mathcal{P}_n = \langle x_1, \ldots, x_n \rangle$ and $f(\mathcal{P}_n)$
  Select $x_{n+1} \in [\ell, u]$ using some strategy and using info from $\mathcal{P}_n, f(\mathcal{P}_n)$ and $s_n(x)$
  Evaluate $f(x_{n+1})$ and set $n \leftarrow n+1$
end while
Return $x_n^* \leftarrow \text{argmin}_{\widetilde{x} \in S_n} f(\widetilde{x})$

In Algorithm 1, the initial sample points $\mathcal{P}_{n_0}$ normally come from a space-filling design such as a Latin hypercube design (LHD). The surrogate $s_n(x)$ is an approximation of $f(x)$ and it may be an interpolating model (i.e., $s_n(x_i) = f(x_i)$ for all $i = 1, \ldots, n$) or a non-interpolating model. Two of the most widely used surrogate models are Gaussian Process (GP) models [13] and Radial Basis Function (RBF) models [12], which are both interpolating models. Because evaluating f is expensive, we often carefully select the next sample point x_{n+1} using some scheme that is either expected (though not guaranteed) to yield a better function value or improve the quality of the current surrogate, or both. The strategy to select the next sample point x_{n+1} is called an *infill criterion* and this usually

involves optimizing an *acquisition function*. Surrogate optimization methods are typically distinguished by the type of surrogate for f and by the infill criterion or acquisition function used to select the next sample point x_{n+1}.

The most popular surrogate optimization approach at present is Bayesian optimization [2,7,9], which uses Gaussian Process surrogates [13] together with some widely used acquisition functions such as the expected improvement and lower confidence bound criteria. A tutorial on this method can be found in [6].

An alternative to Bayesian optimization includes RBF methods such as the RBF method by Gutmann [8] and the Stochastic RBF method [15]. RBF methods, particularly variants of the latter, have been quite successful on high-dimensional test problems and application problems involving hundreds of decision variables (e.g., [14]).

3 Using ChatGPT to Generate Matlab Codes for Expensive Black-Box Optimization

Below, we present some prompts used to generate Matlab codes for expensive black-box optimization. A total of four Matlab codes were generated by the AI, two implement Bayesian optimization while the other two implement RBF methods. In all cases, the codes produced were well-structured and they include proper documentation and references, which are mostly accurate. Because of space limitations, only one of each type of method are included in the Appendix.

First, the following prompt was used on GPT-5.2 to generate a Matlab code that is suitable for expensive black-box global optimization where the variables are subject only to bound constraints.

Please provide a Matlab function that searches for a global minimum point of a computationally expensive black-box objective function subject only to bound constraints. Use an algorithm that attempts to find the global minimum using only a relatively small number of objective function evaluations and without using the gradient of the objective function. The inputs to the Matlab function include the objective function to minimize, the bounds on the decision variables, and the computational budget in terms of number of function evaluations. If using algorithms and ideas in the literature, please provide proper citations of books or articles. Please also demonstrate how to use this Matlab function on three very different widely used test functions whose dimensions are 4, 7 and 10.

The AI generated a Matlab code called bbopt_global_min.m that implements standard Bayesian optimization with the Expected Improvement (EI) criterion. This outcome was expected since Bayesian optimization is the currently the most popular algorithm for expensive black-box optimization. Moreover, the EI criterion is the classic acquisition function associated with Bayesian optimization that was introduced in the paper on Efficient Global Optimization (EGO) [9]. However, unlike the original implementation of EGO and other Bayesian optimization codes, the AI-generated code simply maximizes the EI criterion over random candidate points instead of doing a more thorough search. This

idea of optimizing an acquisition function over random candidate points is similar to that used by the Stochastic RBF method [15] and it results in much faster code, especially when using a small number of random candidate points relative to the dimension of the problem. In this case, the AI-generated code uses a default of 5000 candidate points.

Next, GPT-5.2 was given the following prompt to generate an alternative surrogate optimization method that uses RBF surrogates:

Are you able to provide a similar Matlab function for computationally expensive bound-constrained black-box global optimization that uses radial basis function surrogates instead of Gaussian process surrogates with the Bayesian optimization idea? Again, please provide proper citations of books or articles.

The AI then generated a code called rbfopt_global_min.m that uses RBF interpolation and ideas from the Stochastic RBF paper [15] such as the use of the distance-based criterion. However, the acquisition function it used was not the same as that in the original paper. Instead, the structure is more similar to that of the Lower Confidence Bound idea that is also used in Bayesian optimization except that the uncertainty is replace by a distance-based proxy where points that are farther away from current sample points are associated with higher uncertainty in the surrogate approximation. Hence, GPT-5.2 seems to have generated a novel and reasonable surrogate optimization method that is a hybrid of ideas used in RBF methods and Bayesian optimization.

At the end of the response, the AI also offered to generate a code that is closer to the Stochastic RBF algorithm by Regis and Shoemaker [15]. Hence, the following prompt was then given to the AI:

Please provide a closer-to-Regis/Shoemaker variant of the code that uses RBF with the same input-output structure.

This resulted in another code called rbfopt_global_min_srbf.m that was indeed more similar to the Stochastic RBF method called MLMSRBF [15] except that the default weights for the RBF and distance criteria in the acquisition function were different from the one used in the original MLMSRBF algorithm. Instead, they were taken from an extension of the algorithm to constrained black-box optimization [14]. A copy of this code is included in the Appendix. To come up with an even closer implementation of MLMSRBF, the AI-generated code was also run with the same weights used as in the original MLMSRBF algorithm. This gave rise to two variants of this code: RBFOPT-SRBF-A (uses the default weights provided by the AI) and RBFOPT-SRBF-B (uses the same weights as MLMSRBF).

Finally, the BBOPT algorithm that uses Bayesian optimization yielded relatively poor results on the higher dimensional test problems, so the AI was prompted to generate a variant of Bayesian optimization that runs more quickly and that is more suitable for higher dimensional problems:

Can you create a variant of the bbopt_global_min.m code that runs more quickly and that is more suitable for higher dimensional optimization problems with 10 or more decision variables?

GPT-5.2 then generated the code bbopt_global_min_fast.m, which is a Matlab implementation of the TuRBO (Trust Region Bayesian Optimization) algorithm that was designed for higher dimensional problems. A copy of this code is included in the Appendix. The original TuRBO algorithm is implemented in Python [5], so the AI-generated code is convenient for users whose simulation code for the black-box function runs on Matlab. TuRBO employs a trust region approach with a Gaussian Process model built using the k nearest sampled points from the trust region center. The next sample point is obtained by minimizing a lower confidence bound (LCB) acquisition function within the current trust region. This approach is meant to promote local search, which facilitates progress on high dimensional problems.

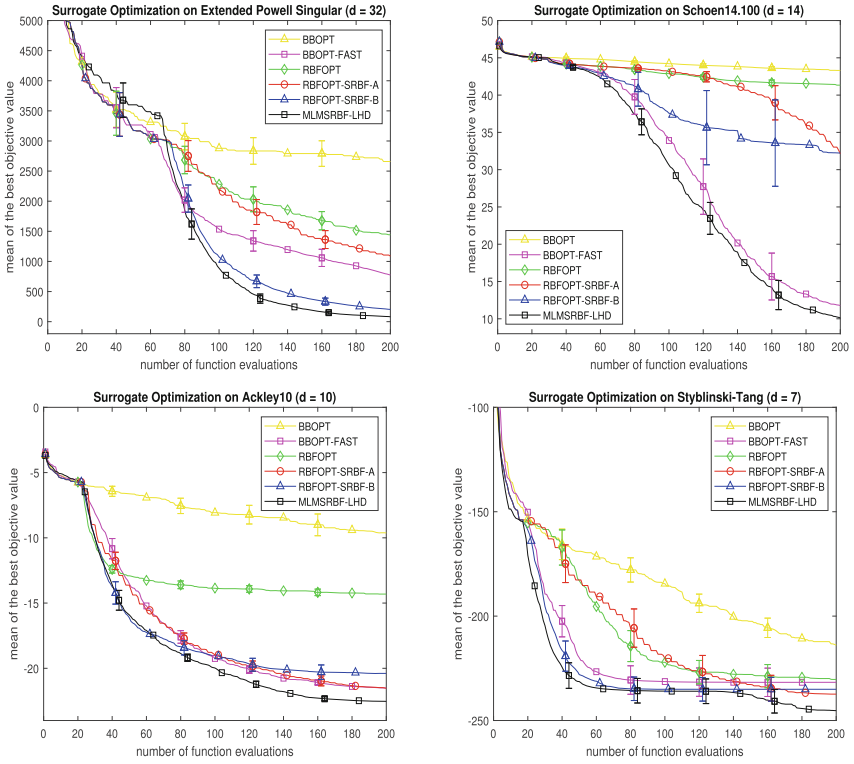


Fig. 1. Mean of the best objective function values (over 30 trials) obtained by the AI-generated surrogate optimization codes and the MLMSRBF-LHD algorithm on the Extended Powell Singular (top left), Schoen14.100 (top right), Ackley10 (bottom left) and Styblinski-Tang (bottom right) test problems. In all plots, the error bars correspond to 95% t confidence intervals for the mean.

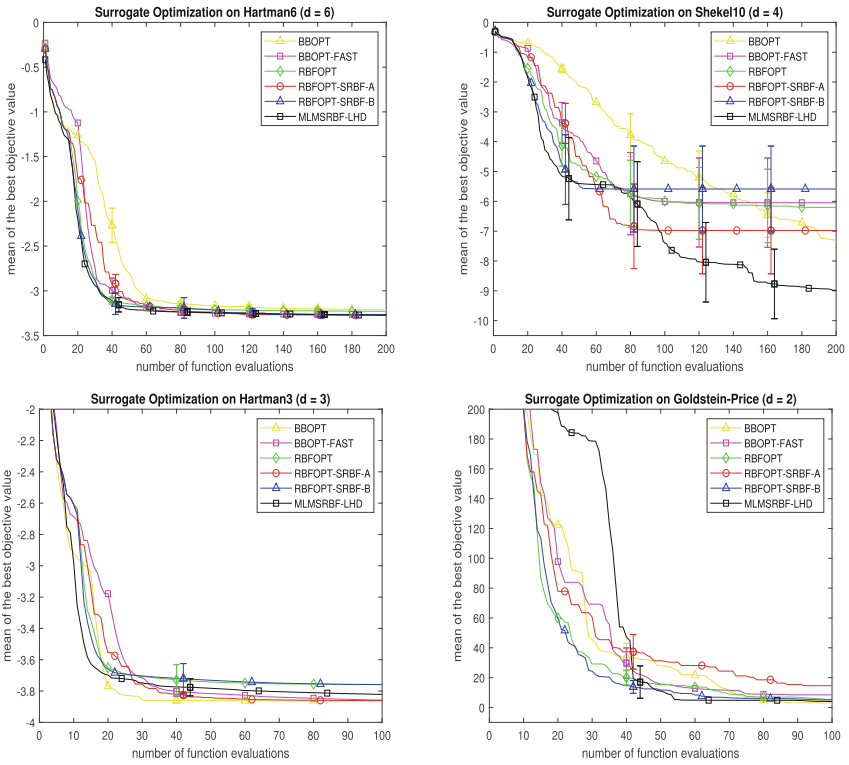


Fig. 2. Mean of the best objective function values (over 30 trials) obtained by the AI-generated surrogate optimization codes and the MLMSRBF-LHD algorithm on the Hartman6 (top left), Shekel10 (top right), Hartman3 (bottom left), and Goldstein-Price (bottom right) test problems. In all plots, the error bars correspond to 95% t confidence intervals for the mean.

4 Experimental Setup

We compare Matlab codes generated by GPT-5.2 with an original human-generated Matlab code of the Multistart Local Metric Stochastic RBF (MLMSRBF) [15] on eight well-known test problems and on a 36-dimensional groundwater bioremediation application described in [22]. Groundwater bioremediation is the process of promoting the growth of microorganisms that are then used to break down hazardous contaminants in the groundwater. Each of the AI-generated codes BBOPT, BBOPT-FAST, RBFOPT were run using the default settings. The AI-generated code RBFOPT-SRBF, which is an AI implementation that resembles MLMSRBF was run with two variants: RBFOPT-SRBF-A uses the default settings used by the AI while RBFOPT-SRBF-B is the exactly the same except it uses the same weights for the RBF and distance criteria as in MLMSRBF. The human-generated MLMSRBF code that was run comes from

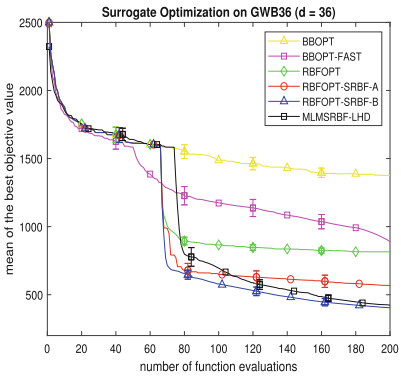


Fig. 3. Mean of the best objective function values (over 30 trials) obtained by the AI-generated surrogate optimization codes and the MLMSRBF-LHD algorithm on the 36-dimensional groundwater bioremediation application GWB36. In both plots, the error bars correspond to 95% t confidence intervals for the mean.

the GitHub repository https://github.com/INFORMSJoC/1060.0182 and used mostly the default settings in [15] except it used an approximate maximin Latin hypercube design (LHD) of size $2(d+1)$ instead of a symmetric LHD. The test problems and their dimensions are the Extended Powell Singular ($d = 32$) [11], Ackley10 ($d = 10$) [1], Styblinski-Tang ($d = 7$) [18] and five of the test problems used in [15], namely, Schoen14.100 ($d = 14$), Hartman6 ($d = 6$), Shekel10 ($d = 4$), Hartman3 ($d = 3$) and Goldstein-Price ($d = 2$).

To obtain average case performance, each of the AI-generated codes and the original MLMSRBF code were run for 30 trials on each of the test problems and on the groundwater bioremediation application.

5 Results and Discussion

Figures 1-2 show the plots of the mean best objective function value (over 30 trials) obtained by the AI-generated surrogate optimization codes and the MLMSRBF-LHD algorithm on the eight synthetic test problems. The error bars correspond to 95% t confidence intervals for the mean. Moreover, Fig. 3 show the corresponding plots on the 36-D groundwater bioremediation application.

The plots show that the original MLMSRBF-LHD algorithm generally performs better than the AI-generated surrogate optimization codes on five of the synthetic test problems. Moreover, its performance is comparable with that of the AI-generated surrogate optimization codes on Hartman6 and Hartman3. However, the RBFOPT-SRBF-B code, which GPT-5.2 designed to be more similar to the MLMSRBF-LHD algorithm, was generally better than the latter on the 36-D groundwater bioremediation application and on the 2-D Goldstein-Price problem. Moreover, the performance of RBFOPT-SRBF-B is close to that of

MLMSRBF-LHD on the 32-D Extended Powell Singular, the 7-D Styblinski-Tang, Hartman6 and Hartman3. The BBOPT-FAST code has comparable performance to MLMSRBF-LHD on the Goldstein-Price, Hartman3 and Hartman6, and its mean objective function value is not too far from that of the MLMSRBF-LHD on Schoen14.100, Ackley10, and Styblinski-Tang. Overall, the performances of RBFOPT-SRBF-B and BBOPT-FAST are close to that of MLMSRBF-LHD on the test problems. Additionally, the somewhat novel RBF-OPT performs better than BBOPT-FAST on GWB36 and Goldstein-Price and has comparable performance to BBOPT-FAST on Hartman6, Shekel10 and Hartman3.

6 Summary and Future Work

This study compared the Matlab codes for surrogate optimization that were generated by ChatGPT 5.2 with a human-generated original Matlab code that implements Multistart Local Metric Stochastic RBF (MLMSRBF), which is an established RBF method for expensive black-box optimization. The codes were compared on eight well-known optimization test problems with up to 32 dimensions and on a 36-dimensional groundwater bioremediation application. The codes generated by GPT-5.2 are well-organized and well-documented and they implement standard surrogate optimization approaches. In the case of the AI-generated RBFOPT code, GPT-5.2 created a somewhat novel RBF algorithm that minimizes a lower confidence bound acquisition function that is commonly used in Bayesian optimization. The numerical results show that the AI-generated BBOPT-FAST and RBFOPT-SRBF-B codes achieved comparable performance to the human-generated MLMSRBF code on the test problems. Moreover, on the groundwater application, RBFOPT-SRBF-B performed better than the human-generated MLMSRBF code while RBFOPT-SRBF-A achieved comparable performance with MLMSRBF. BBOPT-FAST is an AI-generated TuRBO-style Bayesian optimization code while RBFOPT-SRBF-B is an AI-generated Stochastic RBF-style code that uses weights for the RBF and distance criteria that are similar to MLMSRBF. These results indicate that AI tools can be used to generate high quality Matlab codes for surrogate-based expensive black-box optimization. They also suggest great potential for AI tools to assist in developing novel optimization algorithms by combining ideas from multiple methods. Future work will experiment with using the AI to generate promising infill criteria for expensive black-box optimization. We also examine AI-generated Matlab codes for constrained optimization and multi-objective optimization and codes in other languages such as R or Python.

Acknowledgments. This study used ChatGPT Enterprise with GPT-5.2 from the Wharton School of the University of Pennsylvania.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

A Appendix

Below is the Matlab code for the AI-generated BBOPT-FAST algorithm (generated by GPT-5.2), which is the fast high-dimensional variant of the AI-generated BBOPT algorithm that implements the TuRBO-style Bayesian optimization:

```
function [xbest, fbest, history] = bbopt_global_min_fast(objfun, lb, ub, budget, opts)
%BBOPT_GLOBAL_MIN_FAST  Faster BO variant for d >= 10 using a TuRBO-style trust region.
%
%   [xbest, fbest, history] = bbopt_global_min_fast(objfun, lb, ub, budget, opts)
%
% Inputs/Outputs match bbopt_global_min.m.
%
% Algorithm: Trust-Region Bayesian Optimization (TuRBO-style):
% - Maintain a trust region (TR) around current best.
% - Fit a *local* GP on k-nearest sampled points to the TR center.
% - Propose next point by minimizing LCB: mu(x) - beta * sigma(x) over candidates in TR.
% - Expand TR on repeated successes; shrink on failures.
%
% Reference:
% - Eriksson et al., "Scalable Global Optimization via Local Bayesian Optimization", NeurIPS 2019.

    if nargin < 5, opts = struct(); end
    lb = lb(:)'; ub = ub(:)';
    d = numel(lb);
    assert(numel(ub) == d && all(ub > lb), 'Invalid bounds.');
    assert(budget >= 1 && budget == floor(budget), 'budget must be a positive integer');

    % ----- Defaults tuned for higher-dimensional (>=10) -----
    if ~isfield(opts,'seed'), opts.seed = []; end
    if ~isfield(opts,'verbose'), opts.verbose = true; end

    % Initial design (keep modest)
    if ~isfield(opts,'nInit')
        opts.nInit = min(max(2*d, 20), floor(budget/4));
    end

    % Candidate count per iteration (cheap; mostly vectorized)
    if ~isfield(opts,'nCand'), opts.nCand = max(4000, 400*d); end

    % Local GP: use only k nearest points
    if ~isfield(opts,'kLocal'), opts.kLocal = min(60, max(20, 5*d)); end

    % GP kernel settings (fixed/heuristic for speed)
    if ~isfield(opts,'kernel'), opts.kernel = 'matern52'; end % 'matern52' or 'se'
    if ~isfield(opts,'ell'),    opts.ell = 0.25; end         % lengthscale in normalized space (scalar)
    if ~isfield(opts,'sn'),     opts.sn = 1e-6; end          % nugget/noise
    if ~isfield(opts,'sf'),     opts.sf = 1.0; end           % signal std

    % Acquisition: LCB coefficient schedule
    if ~isfield(opts,'beta0'),  opts.beta0 = 2.0; end        % initial exploration
    if ~isfield(opts,'betaMin'),opts.betaMin = 0.5; end      % later exploration

    % Trust region parameters (side length in normalized space)
    if ~isfield(opts,'tr0'),     opts.tr0 = 0.8; end
    if ~isfield(opts,'trMin'),   opts.trMin = 0.05; end
    if ~isfield(opts,'trMax'),   opts.trMax = 1.0; end
    if ~isfield(opts,'succTol'), opts.succTol = 3; end
    if ~isfield(opts,'failTol'), opts.failTol = 6; end
    if ~isfield(opts,'trGrow'),  opts.trGrow = 1.5; end
    if ~isfield(opts,'trShrink'),opts.trShrink = 0.7; end
```

```
if ~isempty(opts.seed), rng(opts.seed); end

% Normalize domain to [0,1]^d for internal work
toUnit  = @(x) (x - lb) ./ (ub - lb);
fromUnit = @(u) lb + u .* (ub - lb);

nInit = min(opts.nInit, budget);

% --- Initial design (Sobol if available; else LHS) ---
U = initialDesign(nInit, d);  % in [0,1]^d
X = fromUnit(U);
f = zeros(nInit,1);
for i = 1:nInit
    f(i) = objfun(X(i,:));
end

[fbest, idx] = min(f);
xbest = X(idx,:);
ubest = toUnit(xbest);

% Trust region state
tr = min(max(opts.tr0, opts.trMin), opts.trMax);
nSucc = 0; nFail = 0;

history.X = X;
history.f = f;
history.xbest = xbest;
history.fbest = fbest;
history.tr = tr;
history.beta = [];
history.lcbMin = [];

if opts.verbose
    fprintf('BBOPT_FAST(TuRBO): init n=%d, best f=%.6g\n', nInit, fbest);
end

for t = (nInit+1):budget
    % LCB beta schedule: decrease exploration over time
    frac = (t-1)/(max(budget-1,1));
    beta = max(opts.betaMin, opts.beta0 * (1 - 0.8*frac));

    % Build candidate set in trust region around ubest
    Ucand = sampleInTrustRegion(ubest, tr, opts.nCand);

    % Select local data (k nearest in normalized space)
    Uall = toUnit(X);
    idxLocal = knnIndices(Uall, ubest, min(opts.kLocal, size(Uall,1)));
    Uloc = Uall(idxLocal,:);
    yloc = f(idxLocal);

    % Standardize y locally for numerical stability
    ymean = mean(yloc);
    ystd = std(yloc); if ystd < 1e-12, ystd = 1; end
    ys = (yloc - ymean) / ystd;

    % GP predict (vectorized)
    [mu_s, s2_s] = gpPredictLocal(Uloc, ys, Ucand, opts);

    mu = ymean + ystd * mu_s;
    s  = sqrt(max(0, (ystd^2) * s2_s));

    lcb = mu - beta * s;
    [lcbMin, j] = min(lcb);
    unext = Ucand(j,:);
```

```
        xnext = fromUnit(unext);

        fnext = objfun(xnext);

        % Append
        X = [X; xnext];
        f = [f; fnext];

        % Update incumbent and TR logic
        if fnext < fbest
            fbest = fnext;
            xbest = xnext;
            ubest = unext;
            nSucc = nSucc + 1;
            nFail = 0;
        else
            nFail = nFail + 1;
            nSucc = 0;
        end

        if nSucc >= opts.succTol
            tr = min(opts.trMax, tr * opts.trGrow);
            nSucc = 0;
        elseif nFail >= opts.failTol
            tr = max(opts.trMin, tr * opts.trShrink);
            nFail = 0;
        end

        % Save history
        history.X = X;
        history.f = f;
        history.xbest = xbest;
        history.fbest = fbest;
        history.tr(end+1,1) = tr;
        history.beta(end+1,1) = beta;
        history.lcbMin(end+1,1) = lcbMin;

        if opts.verbose
            fprintf('BBOPT_FAST: eval %3d/%3d | fnext=%.6g | best=%.6g | TR=%.3g | beta=%.3g\n',...
                t, budget, fnext, fbest, tr, beta);
        end
    end
end

% ============================ Helpers ============================

function U = initialDesign(n, d)
% Try Sobol for better space filling; else LHS.
    try
        p = sobolset(d, 'Skip', 1e3, 'Leap', 1e2);
        p = scramble(p,'MatousekAffineOwen');
        U = net(p, n);
    catch
        U = lhsdesign(n, d, 'criterion','maximin','iterations',30);
    end
end

function Ucand = sampleInTrustRegion(uc, tr, nCand)
% Sample uniformly in a hyper-rectangle trust region centered at uc in [0,1]^d.
    d = numel(uc);
    half = 0.5 * tr;
    lo = max(0, uc - half);
    hi = min(1, uc + half);
    Ucand = rand(nCand, d) .* (hi - lo) + lo;
```

```
end

function idx = knnIndices(U, u0, k)
% Return indices of k nearest points in Euclidean distance in [0,1]^d.
    dif = U - u0;
    dist2 = sum(dif.^2, 2);
    [~, order] = sort(dist2, 'ascend');
    idx = order(1:k);
end

function [mu, s2] = gpPredictLocal(X, y, Xstar, opts)
% Lightweight GP posterior with fixed hyperparams (no optimization).
% X, Xstar in [0,1]^d, y standardized.
% Kernel: Matern 5/2 (default) or SE. Uses scalar lengthscale for speed.
    n = size(X,1);
    sn2 = opts.sn^2;
    sf2 = opts.sf^2;
    ell = opts.ell;

    % Compute K (n x n)
    D = pdist2_fast(X, X) / ell;
    K = sf2 * kernelFromScaledDist(D, opts.kernel);
    K = K + (sn2 + 1e-10) * eye(n);

    L = chol(K, 'lower');
    alpha = L'\(L\y);

    % Cross-cov (n x m)
    Ds = pdist2_fast(X, Xstar) / ell;
    Kxs = sf2 * kernelFromScaledDist(Ds, opts.kernel);

    mu = Kxs' * alpha;

    v = L \ Kxs;

    % Prior variance k(x*,x*) = sf2 * k(0) where k(0)=1 for these kernels
    s2 = sf2 - sum(v.^2, 1)';
    s2 = max(s2, 0);
end

function K = kernelFromScaledDist(D, kernel)
% D is scaled Euclidean distance (already / ell)
    switch lower(kernel)
        case 'matern52'
            % k(r) = (1 + sqrt(5)r + 5r^2/3) exp(-sqrt(5)r)
            r = D;
            a = sqrt(5) * r;
            K = (1 + a + (5/3)*(r.^2)) .* exp(-a);
        case 'se'
            % k(r) = exp(-0.5 r^2)
            K = exp(-0.5 * (D.^2));
        otherwise
            error('Unknown kernel: use ''matern52'' or ''se''.');
    end
end

function D = pdist2_fast(A, B)
% Fast Euclidean distance (like pdist2) without toolbox dependency.
% Returns size(A,1) x size(B,1).
    a2 = sum(A.^2, 2);
    b2 = sum(B.^2, 2);
    D2 = a2 + b2' - 2*(A*B');
    D2 = max(D2, 0);
    D = sqrt(D2);
```

```
end
```

Below is the Matlab code for the RBFOPT-SRBF algorithm generated by GPT-5.2:

```
function [xbest, fbest, history] = rbfopt_global_min_srbf(objfun, lb, ub, budget, opts)
%RBFOPT_GLOBAL_MIN_SRBF  Stochastic RBF (SRBF)-style expensive global optimization
% under bound constraints (Regis & Shoemaker-like).
%
%   [xbest, fbest, history] = rbfopt_global_min_srbf(objfun, lb, ub, budget, opts)
%
% Inputs
%   objfun : handle, f = objfun(x), x is 1xd
%   lb, ub : 1xd bounds
%   budget : number of function evaluations allowed
%   opts   : optional struct
%       .nInit        initial design size (default: min(max(2*d+1,10), floor(budget/3)))
%       .nCand        total # candidates per iteration (default: 5000)
%       .localFrac    fraction of candidates generated locally near incumbent (default: 0.7)
%       .perturbProb  probability each coordinate is perturbed in local candidates (default: min(0.2, 5/d))
%       .sigma0       initial local perturbation scale as fraction of (ub-lb) (default: 0.2)
%       .sigmaMin     minimum sigma fraction (default: 0.02)
%       .sigmaMax     maximum sigma fraction (default: 0.5)
%       .sigmaShrink  factor to shrink sigma after failures (default: 0.7)
%       .sigmaGrow    factor to grow sigma after successes (default: 1.2)
%       .failTol      consecutive failures before shrinking sigma (default: 5)
%       .succTol      consecutive successes before growing sigma (default: 3)
%       .weightCycle  vector of weights for predicted objective in merit (default: [0.3 0.5 0.7 0.9])
%                     distance weight is (1-w). Cycled over iterations (SRBF-like).
%       .rbf          'cubic' (default) or 'thinplate'
%       .reg          RBF system regularization (default: 1e-10)
%       .seed         RNG seed (default: [])
%       .verbose      true/false (default: true)
%
% Outputs
%   xbest, fbest : incumbent best solution/value
%   history      : struct with fields X, f, xbest, fbest, sigma, weight, meritBest
%
% Notes
% - This is an SRBF-style algorithm: RBF surrogate + local/global random candidates
%   + merit function combining predicted value and distance-to-samples.
%
% References:
% - Regis & Shoemaker (2007) INFORMS J. on Computing, SRBF method.
% - Regis & Shoemaker (2009) INFORMS J. on Computing, parallel SRBF variants.
% - Wendland (2005), Powell (1987) RBF interpolation foundations.

    if nargin < 5, opts = struct(); end
    lb = lb(:)'; ub = ub(:)';
    d = numel(lb);
    assert(numel(ub)==d && all(ub>lb), 'Invalid bounds');
    assert(budget>=1 && budget==floor(budget), 'budget must be a positive integer');

    % Defaults
    if ~isfield(opts,'nInit'),       opts.nInit = min(max(2*d+1,10), floor(budget/3)); end
    if ~isfield(opts,'nCand'),       opts.nCand = 5000; end
    if ~isfield(opts,'localFrac'),   opts.localFrac = 0.7; end
    if ~isfield(opts,'perturbProb'), opts.perturbProb = min(0.2, 5/d); end
    if ~isfield(opts,'sigma0'),      opts.sigma0 = 0.2; end
    if ~isfield(opts,'sigmaMin'),    opts.sigmaMin = 0.02; end
    if ~isfield(opts,'sigmaMax'),    opts.sigmaMax = 0.5; end
    if ~isfield(opts,'sigmaShrink'), opts.sigmaShrink = 0.7; end
    if ~isfield(opts,'sigmaGrow'),   opts.sigmaGrow = 1.2; end
```

```
if ~isfield(opts,'failTol'),     opts.failTol = 5; end
if ~isfield(opts,'succTol'),     opts.succTol = 3; end
if ~isfield(opts,'weightCycle'), opts.weightCycle = [0.3 0.5 0.7 0.9]; end
if ~isfield(opts,'rbf'),         opts.rbf = 'cubic'; end
if ~isfield(opts,'reg'),         opts.reg = 1e-10; end
if ~isfield(opts,'seed'),        opts.seed = []; end
if ~isfield(opts,'verbose'),     opts.verbose = true; end

if ~isempty(opts.seed), rng(opts.seed); end

nInit = min(opts.nInit, budget);
nCand = max(1000, opts.nCand);

% --- Initial design (LHS)
X = lhsdesign(nInit, d, 'criterion','maximin', 'iterations', 50);
X = bsxfun(@plus, lb, bsxfun(@times, X, (ub-lb)));
f = zeros(nInit,1);
for i=1:nInit
    f(i) = objfun(X(i,:));
end

[fbest, idx] = min(f);
xbest = X(idx,:);

% Local step size (fraction of range)
sigmaFrac = min(max(opts.sigma0, opts.sigmaMin), opts.sigmaMax);
nFail = 0; nSucc = 0;

history.X = X;
history.f = f;
history.xbest = xbest;
history.fbest = fbest;
history.sigma = [];
history.weight = [];
history.meritBest = [];

if opts.verbose
    fprintf('SRBF-RBFOPT: init n=%d, best f=%.6g\n', nInit, fbest);
end

wCycle = opts.weightCycle(:)';
nW = numel(wCycle);

% --- Main loop
for t = (nInit+1):budget
    % Fit RBF surrogate
    model = fitRBF(X, f, opts.rbf, opts.reg);

    % Choose merit weight (cycle)
    w = wCycle( mod(t-1, nW) + 1 );      % weight on predicted objective
    w = min(max(w, 0), 1);
    wd = 1 - w;                          % weight on distance term

    % Generate candidates: local + global
    nLocal = round(opts.localFrac * nCand);
    nGlobal = nCand - nLocal;

    Xlocal = generateLocalCandidates(xbest, lb, ub, nLocal, sigmaFrac, opts.perturbProb);
    Xglob  = rand(nGlobal, d);
    Xglob  = bsxfun(@plus, lb, bsxfun(@times, Xglob, (ub-lb)));
    Xcand  = [Xlocal; Xglob];

    % Predict surrogate value at candidates
    fhat = rbfPredict(model, Xcand);
```

```
        % Distance-to-existing points (encourage exploring sparse regions)
        dmin = nearestDistance(Xcand, X);

        % Normalize both criteria to [0,1] (SRBF-style merit uses scaled quantities)
        fhatN = normalize01(fhat);
        dN    = normalize01(dmin);

        % Merit for minimization: smaller is better
        % Predicted objective term: fhatN
        % Exploration term: want LARGE distance => subtract dN
        merit = w * fhatN - wd * dN;

        [meritBest, j] = min(merit);
        xnext = Xcand(j,:);

        % Evaluate expensive objective
        fnext = objfun(xnext);

        % Append
        X = [X; xnext];
        f = [f; fnext];

        improved = (fnext < fbest);
        if improved
            fbest = fnext;
            xbest = xnext;
            nSucc = nSucc + 1;
            nFail = 0;
        else
            nFail = nFail + 1;
            nSucc = 0;
        end

        % SRBF-style step-size adaptation
        if nFail >= opts.failTol
            sigmaFrac = max(opts.sigmaMin, sigmaFrac * opts.sigmaShrink);
            nFail = 0;
        elseif nSucc >= opts.succTol
            sigmaFrac = min(opts.sigmaMax, sigmaFrac * opts.sigmaGrow);
            nSucc = 0;
        end

        history.X = X;
        history.f = f;
        history.xbest = xbest;
        history.fbest = fbest;
        history.sigma(end+1,1) = sigmaFrac;
        history.weight(end+1,1) = w;
        history.meritBest(end+1,1) = meritBest;

        if opts.verbose
            fprintf('SRBF-RBFOPT: eval %3d/%3d | fnext=%.6g | best=%.6g | w=%.2f | sigma=%.3g\n', ...
                t, budget, fnext, fbest, w, sigmaFrac);
        end
    end
end

% ======================== Helper functions ========================

function Xcand = generateLocalCandidates(xbest, lb, ub, n, sigmaFrac, perturbProb)
% Generate local candidates by perturbing xbest coordinate-wise with probability perturbProb.
% Perturbations are Gaussian with std = sigmaFrac*(ub-lb).
    d = numel(lb);
```

```
    if n <= 0
        Xcand = zeros(0,d);
        return;
    end

    range = (ub - lb);
    sig = sigmaFrac * range;

    Xcand = repmat(xbest, n, 1);

    % For each candidate, choose coordinates to perturb
    mask = rand(n,d) < perturbProb;

    % Ensure at least one coordinate perturbed per candidate
    rowsNoPert = find(~any(mask,2));
    if ~isempty(rowsNoPert)
        idx = randi(d, numel(rowsNoPert), 1);
        mask(sub2ind([n d], rowsNoPert, idx)) = true;
    end

    Z = randn(n,d);
    Xcand(mask) = Xcand(mask) + Z(mask).*repmat(sig, n, 1)(mask);

    % Project to bounds
    Xcand = min(max(Xcand, lb), ub);
end

function z = normalize01(v)
    v = v(:);
    vmin = min(v); vmax = max(v);
    if abs(vmax - vmin) < 1e-12
        z = zeros(size(v));
    else
        z = (v - vmin) / (vmax - vmin);
    end
end

function model = fitRBF(X, y, rbfType, reg)
% RBF interpolant with affine tail:
% s(x) = sum_i lambda_i * phi(||x-x_i||) + [1 x]^T beta
    [n, d] = size(X);
    y = y(:);

    R = pdist2(X, X);
    Phi = phiRBF(R, rbfType);
    P = [ones(n,1), X];

    A = [Phi + reg*eye(n), P;
         P', zeros(d+1,d+1)];
    b = [y; zeros(d+1,1)];

    coeff = A \ b;
    model.X = X;
    model.lambda = coeff(1:n);
    model.beta   = coeff(n+1:end);
    model.rbfType = rbfType;
end

function yhat = rbfPredict(model, Xq)
    X = model.X;
    R = pdist2(Xq, X);
    Phi = phiRBF(R, model.rbfType);
    Pq = [ones(size(Xq,1),1), Xq];
    yhat = Phi * model.lambda + Pq * model.beta;
```

```
end

function Phi = phiRBF(R, rbfType)
    switch lower(rbfType)
        case 'cubic'
            Phi = R.^3;
        case 'thinplate'
            Phi = R.^2 .* log(R + eps); % phi(0)=0 automatically via eps
        otherwise
            error('Unknown rbfType. Use ''cubic'' or ''thinplate''.');
    end
end

function dn = nearestDistance(Xcand, X)
% Distance from each candidate to nearest sampled point
    D = pdist2(Xcand, X);
    dn = min(D, [], 2);
end
```

References

1. Ackley, D.H.: A Connectionist Machine for Genetic Hillclimbing. Kluwer Academic Publishers, Boston, MA (1987)
2. Archetti, F., Candelieri, A.: Bayesian optimization and data science (Vol. 849). Springer, Cham (2019)
3. Bagheri, S., Konen, W., Emmerich, M., Bäck, T.: Self-adjusting parameter control for surrogate-assisted constrained optimization under limited budgets. Appl. Soft Comput. **61**, 377–393 (2017)
4. Eriksson, D., Bindel, D., Shoemaker, C.A.: pySOT and POAP: An event-driven asynchronous framework for surrogate optimization. arXiv preprint arXiv:1908.00420 (2019)
5. Eriksson, D., Pearce, M., Gardner, J., Turner, R.D., Poloczek, M.: Scalable global optimization via local Bayesian optimization. Adv. Neural Inform. Process. Syst. **32** (2019)
6. Frazier, P.I.: A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811 (2018)
7. Garnett, R.: Bayesian optimization. Cambridge University Press (2023)
8. Gutmann, H.-M.: A radial basis function method for global optimization. J. Global Optim. **19**(3), 201–227 (2001)
9. Jones, D., Schonlau, M., Welch, W.: Efficient global optimization of expensive black-box functions. J. Global Optim. **13**(4), 455–492 (1998)
10. Luzzi, M., Guerriero, F., Maratea, M., Greco, G., Garofalo, M.: Chatgpt and operations research: evaluation on the shortest path problem. Soft Comput. **29**, 1407–1418 (2025). https://doi.org/10.1007/s00500-025-10505-2
11. Moré, J., Garbow, B., Hillstrom, K.: Testing unconstrained optimization software. ACM Trans. Math. Softw. **7**, 17–41 (1981)
12. Powell, M.J.D.: The theory of radial basis function approximation in 1990. In: Light, W. (ed.) Advances in Numerical Analysis, Volume 2: Wavelets, Subdivision Algorithms and Radial Basis Functions, pp. 105–210. Oxford University Press, Oxford, U.K. (1992)
13. Rasmussen, C. E.: Gaussian Processes for Machine Learning. The MIT Press (2006)

14. Regis, R.G.: Stochastic radial basis function algorithms for large-scale optimization involving expensive black-box objective and constraint functions. Comput. Oper. Res. **38**(5), 837–853 (2011)
15. Regis, R.G., Shoemaker, C.A.: A stochastic radial basis function method for the global optimization of expensive functions. INFORMS J. Comput. **19**(4), 497–509 (2007)
16. Smith, A.E.: Note from the Editor. INFORMS J. Comput. **37**(4), iv–viii (2005). https://doi.org/10.1287/ijoc.2025.ed.v37.n4
17. Sorokovikov, P., Gornov, A., Zarodnyuk, T.: Numerical Investigation of the Swarm Intelligence Algorithm Obtained Using ChatGPT for Univariate Global Optimization. In: Eremeev, A., Khachay, M., Kochetov, Y., Mazalov, V., Pardalos, P. (eds) Mathematical Optimization Theory and Operations Research: Recent Trends. MOTOR 2024. Communications in Computer and Information Science, vol 2239. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73365-9_25 (2024)
18. Styblinski, M.A., Tang, T.: Experiments in nonconvex optimization: stochastic approximation with function smoothing and simulated annealing. Neural Netw. **3**(4), 467–483 (1990)
19. The MathWorks Inc.: Global Optimization Toolbox, Version R2025b. The MathWorks Inc., Natick, Massachusetts, United States (2025)
20. The MathWorks Inc.: Statistics and Machine Learning Toolbox, Version R2025b. The MathWorks Inc., Natick, Massachusetts, United States (2025)
21. Vu, N.G.H., Wang, K., Wang, G.G.: Effective prompting with ChatGPT for problem formulation in engineering optimization. Eng. Optim. **57**(12), 3677–3694 (2025)
22. Yoon, J.-H., Shoemaker, C.A.: Comparison of optimization methods for groundwater bioremediation. J. Water Resour. Plann. Manage. **125**, 54–63 (1999)
23. Zhou, Q., Sheu, J. B.: The use of Generative Artificial Intelligence (GenAI) in operations research: review and future research agenda. J. Oper. Res. Society 1–21. https://doi.org/10.1080/01605682.2025.2561762 (2025)



Heuristic Approaches for Personnel Scheduling in a Call Center with Social Fairness Considerations

Anaís Veloso Silva, Telmo Pinto(✉), and Cláudio Alves

Centro ALGORITMI/LASI, Universidade do Minho, Braga, Portugal
pg34575@uminho.pt, {telmo,claudio}@dps.uminho.pt

Abstract. This work addresses a personnel scheduling problem arising in a call center company operating 24 h a day, 7 d a week, in which personnel scheduling is currently performed manually. The main objective is to develop and implement heuristic methods that can produce high-quality schedules in less time than the current manual process. The proposed approach starts by generating an initial solution through a constructive heuristic, which is then improved using local search procedures. The aim is to assign employees to work shifts while maximizing shift stability for each employee within each planning period. A second local search procedure was also implemented to reduce disparities in the number of non-working days assigned to employees, thus incorporating a social fairness perspective into the scheduling process. The constructive heuristic and both local search methods were tested on a set of instances including real instances derived from a call center, as well as randomly generated instances. The proposed approach was assessed by comparing the initial solutions with those obtained after applying the local search procedures.

Keywords: Personnel scheduling · Heuristics · Local search · Social fairness

1 Introduction

Personnel scheduling problems involve multiple decisions and have a significant impact on both financial performance and customer and employee satisfaction. In order to reduce labour costs, which arise in part from personnel scheduling decisions, and to maximise effectiveness, organisations seek work schedules that cover demand efficiently [3,10]. An inadequate personnel scheduling process may require the hiring of temporary workers, thereby increasing costs, may lead to penalties due to non-compliance with legal constraints, may increase waiting times, and may result in employee and customer dissatisfaction. Therefore, providing flexible schedules that promote employee well-being and customer satisfaction, while simultaneously ensuring an optimised personnel assignment that avoids understaffing or overstaffing and satisfies demand, is essential for organisations.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 266–282, 2027.
https://doi.org/10.1007/978-3-032-30521-3_17

In the context of customer satisfaction, call centers play a central role, as they are primarily responsible for the interaction between organisations and customers and, therefore, constitute an important operational resource. The efficiency of these services is critical, since long waiting times may result in reduced customer satisfaction; accordingly, call centers must keep waiting times as low as possible. On the other hand, in order to maintain low labour costs, overstaffing should be avoided. Personnel scheduling in this sector, where several objectives must be considered and these are often conflicting, is a complex task. For this reason, call center services increasingly rely on decision-support tools to assist in this process. However, in many organisations, both in the service sector and in industry, personnel scheduling is still performed manually.

This paper addresses an optimisation problem arising in the context of personnel scheduling in a call center service company that operates 24 h a day, 7 d a week, and in which the scheduling process is carried out manually on a monthly basis [14]. In this context, the main objective is the development of a heuristic method capable of solving call center personnel scheduling problems in relatively short computational times, by constructing high-quality work schedules including shifts and days off. That is, the aim is to generate schedules that promote employee well-being and satisfy the organisation's requirements in such a way that automatic scheduling becomes preferable to manual scheduling.

The remainder of the paper is organised as follows. Section 2 reviews the state of the art. Section 3 describes the problem under study. Section 4 presents the constructive heuristic approaches. Section 5 introduces the evaluation functions, neighbourhood structures, and local search methods. Section 6 reports the computational results, and Sect. 7 concludes the paper.

2 Related Work

Personnel scheduling problems have been widely studied in both industry and the service sector [4,5]. The service sector, in particular, is labour-intensive and, therefore, effective workforce utilisation and labour cost control are essential to organisational performance [15]. Within the service sector, call centers play a central role in customer acquisition and retention, as they constitute one of the main interfaces between organisations and customers. Since most call center costs are labour-related, personnel scheduling is particularly important to their operational performance [9].

When a customer contacts a call center, the call is answered immediately if an agent is available; otherwise, the customer must wait until one becomes available. Since call arrival rates vary over time and are usually estimated from historical data, forecasts are inherently uncertain. As a result, call centers must balance low waiting times, which are critical to customer satisfaction, with low labour costs, which require avoiding overstaffing. Their objective is therefore typically to provide the required service quality at minimum labour cost [8].

A common performance measure is the service level, defined as the proportion of calls whose waiting time does not exceed a given threshold [2]. In many public-service call centers, minimum service levels are imposed by external authorities,

and failing to meet them may lead to penalties [2]. For these reasons, assigning shifts and days off in call centers is a complex task involving multiple, often conflicting, objectives, which explains the growing use of decision-support tools in this sector [15].

The literature on call center personnel scheduling includes a wide range of Operations Research methods, such as optimisation, queueing theory, and simulation [13]. A common approach consists of first determining the minimum number of employees required in each period of the day to satisfy demand and service-level requirements, and then assigning employees until all required positions are covered.

A method for minimising personnel costs in a single-skill call center is presented in [1], that is, a call center in which employees have the same skill set, have received the same training, and, in principle, provide the same quality of service, under uncertainty and service-level requirements. The implemented algorithm combines simulation and integer linear programming, with both the objective function and the constraints being evaluated by simulation.

In multi-skill call centers, where calls of different types are handled by employees with different skill sets, organisational requirements may be satisfied through several possible skill combinations, making the personnel scheduling process more challenging [2]. For example, [8] presents a hybrid method for solving an integrated personnel and task scheduling problem in call centers and contact centers. This method combines heuristic algorithms with simulation to solve the personnel scheduling and task assignment problems simultaneously.

In [12], the authors propose a simple method for assigning shifts to employees in a multi-skill call center, taking into account the stochastic nature of the call arrival process. This method is limited to call centers with at most five skills. However, it achieves short execution times and high accuracy, producing near-optimal solutions.

In [6], the authors address an integrated personnel and task scheduling problem in a multi-skill call center under several service-quality-related constraints, with the objective of minimising personnel costs. Their approach extends the method proposed in [1] to the multi-skill setting by combining iterative cutting-plane algorithms with integer programming. The authors first present an integer programming formulation for the personnel scheduling problem with service-level constraints over a planning horizon divided into periods and then adapt the ideas of [1] by initially solving a simpler simulation-based version of the problem. Since computational difficulties arise, heuristic procedures are also proposed.

Finally, [7] addresses a personnel scheduling problem in multi-skill call centers handling several types of calls, although the arrival rate of each call type is low. The proposed approach combines a service-level approximation method with a branch-and-bound algorithm to minimise personnel costs subject to minimum service-level requirements.

The proposed approach follows a heuristic optimisation framework combining a constructive heuristic with neighbourhood-based local search procedures,

which are commonly adopted in personnel scheduling problems to generate good-quality solutions within short computational times.

3 Problem Description

The problem under study arises in the context of personnel scheduling in a call center service company. The aim is to assign employees to work shifts over a one-month planning horizon divided into two periods of approximately two weeks each. When the month has an even number of days, both periods have the same length. Otherwise, the second period contains one additional day.

One of the company's objectives is to create some continuity in the assignment of employees to the same shift over a two-week horizon, that is, to maximise the number of times each employee remains assigned to the same work shift for up to two consecutive weeks. The company operates 7 d a week, 24 h a day, and has 48 shifts, starting every 30 min: the first shift begins at 00h00 and the last shift begins at 23h30. A shift corresponds to 9 consecutive hours, including a one-hour break, and can be defined by its starting time.

Employees are distributed across sites in two different cities and organised into teams according to their skill sets. There are 7 teams: 5 operate 24 h a day, 7 d a week; one operates from 06h30 to 23h00 every day; and another operates from 06h30 to 18h30 on weekdays and from 08h00 to 17h00 on weekends. Each employee must have at least 11 h of rest between shifts, may work at most 5 consecutive days per week, and must have at least 2 d off. The company also seeks to avoid overtime, except when required for operational reasons.

Personnel scheduling is currently performed monthly using a spreadsheet. Since both data input and schedule generation are carried out manually, the process requires considerable time and resources and may be difficult to manage efficiently, particularly for large or complex teams. In this context, the purpose of the proposed heuristic approach is to support automatic personnel scheduling while optimising criteria such as minimising shift changes for the same employee within a two-week period. The goal is to produce schedules that, in terms of both quality and required time, are preferable to those obtained manually.

4 Heuristic Approaches

To solve the employee-to-shift scheduling problem described in the previous section, a heuristic optimisation approach was adopted. Since the problem under study is highly constrained and large-scale, the proposed methodology combines a constructive heuristic for generating feasible initial solutions with neighbourhood-based local search procedures for solution improvement. All generated solutions were required to be feasible.

4.1 Solution Representation

In the context of this problem, a feasible solution (S, P) can be represented by a set of employees $S = (S_1, S_2, \ldots, S_{|S|})$. Each employee is associated with the shifts – represented by their starting times – to which they are assigned over the planning horizon of a given month, or with an indication of no assignment. Thus, $P = (P_1, P_2, \ldots, P_{|S|})$, where $P_i = (p_1^i, p_2^i, \ldots, p_d^i)$ for $i = 1, 2, \ldots, |S|$, and p_j^i corresponds to the starting time of the shift assigned to employee i on day j, with $j = 1, 2, \ldots, d$, and d denotes the number of days in the month under consideration.

Figure 1 shows a feasible solution for scheduling 4 employees over a one-week horizon. Each employee is associated with the starting times of their assigned shifts or an indication of no assignment. Employee A is assigned to the 08h00 shift on consecutive working days, with two consecutive days off, marked X. Employee C also works the same shift across the week but has two non-consecutive days off. Employee B has one unassigned day, and Employee D has two unassigned days, both marked with a hyphen. A day is considered unassigned when an employee is neither allocated to any shift nor given a day off. This differs from a non-working day, which encompasses any day on which an employee does not work, whether due to a scheduled day off or the absence of an operational requirement.

S	Employee A	Employee B
P	(08:00, 08:00, 08:30, 08:00, 08:30, X, X)	(10:00, 11:00, 11:00, 10:00, –, 10:00, 10:00)

S	Employee C	Employee D
P	(14:00, 10:00, 10:00, 14:30, X, 14:30, X)	(15:30, 15:30, –, 15:30, –, 15:00, 15:00)

Fig. 1. Example of a solution representation.

4.2 Initial Solution

Initially, a randomly ordered list of available employees, L_{AVAIL}, is created for each team, and its order is kept fixed throughout the first 2-week period of the month. In addition, three other lists are maintained: the list of busy employees, L_{BUSY}, the list of resting employees, L_{REST}, and the list of employees on a day off, L_{OFF}. The organisation's staffing requirements for each shift starting time on each day of the planning horizon are denoted by $MinReq_{h,d}$.

For a given shift, the list of available employees includes those who are not assigned a day off, whose working hours do not overlap with that shift, and who satisfy the minimum 11-hour rest requirement. The list of busy employees contains those whose working hours overlap with the duration of the shift. The

list of resting employees contains those who are still within the 11-hour rest period following a previous shift. Finally, the list of employees on a day off contains those who cannot be assigned to any shift during that period.

On the first day, the list of available employees contains all employees. For each of the 48 shift starting times, employees are assigned in the order in which they appear in L_{AVAIL} until the minimum staffing requirement is met. Whenever an employee is assigned to a shift, that employee is removed from L_{AVAIL} and added to L_{BUSY}. After completing the corresponding 9-hour shift, the employee is moved to L_{REST}.

From the second day onward, the list of available employees is scanned again, and employees are assigned, whenever possible, to the same shift starting time as on the previous day, subject to the staffing requirements. If the required number of employees is not reached, the corresponding shift starting times are recorded in $ShiftsToAssign$, and the remaining available employees are assigned in their current order to those shifts.

After each shift, employees who complete their working period are moved from L_{BUSY} to L_{REST}, and, after the 11-hour rest period, they are returned to L_{AVAIL} or moved to L_{OFF} if they become eligible for days off.

During the first two weeks of the month, employees with an even identification number are assigned 2 consecutive days off, whereas employees with an odd identification number are assigned 2 non-consecutive days off. During the last two weeks, this rule is reversed in order to promote a more equitable assignment of days off over the monthly planning horizon.

For clarity, the constructive heuristic is described in three parts: the first day of the first period, the remaining days within each period, and the first day of the second period. In all algorithms, the main input parameters are the randomly ordered list of available employees by team, L_{AVAIL}, and the daily staffing requirements, $MinReq_{h,d}$. The algorithms also make use of the lists L_{BUSY}, L_{REST}, L_{OFF}, and $ShiftsToAssign$, as well as the auxiliary variables $numWorking$ and $numRequired$. The main functions used are listed in Table 1. In the pseudocode, days are indexed from 0, so that $d = 0$ denotes the first day of the month.

4.3 First Day of the First Period

The procedure for the first day is described in Algorithm 1. The input parameters are the daily staffing requirements, $MinReq_{h,d}$, the randomly ordered list of available employees, L_{AVAIL}, which in this case contains all employees of a given team, and the lists L_{BUSY} and L_{REST}, both initialised as empty. For each shift starting time h, employees are assigned whenever the number of employees currently working, $numWorking$, is smaller than the minimum required number, $MinReq_{h,d}$. Assigned employees are selected in the order in which they appear in L_{AVAIL} and moved to L_{BUSY}. In addition, for each shift starting time, the lists of busy and resting employees are updated so that employees are moved from L_{BUSY} to L_{REST} after completing their shifts, and from L_{REST} back to L_{AVAIL} after the 11-hour rest period.

Table 1. Functions and corresponding functionalities

Function	Meaning
$Size(x)$	Returns the size of list x, i.e., the number of elements in the list.
$SelectEmployee(x)$	Selects the first element of list x.
$AssignShift(j, h)$	Assigns employee j to shift h.
$ReturnShiftEnd(j)$	Returns the shift ending time of employee j.
$ReturnRestEnd(j)$	Returns the time at which employee j completes the 11-hour rest period.
$ReturnDayOffEnd(j)$	Returns the time and day at which employee j completes the assigned day(s) off.
$ReturnPreviousShift(j)$	Returns the starting time of the last shift worked by employee j.
$History(j)$	Returns the record of all shifts worked by employee j.
$ReturnId(j)$	Returns the unique identification number of employee j.
$AssignDaysOff(j, h, d, x)$	When $d \in \left[1, \frac{numDays}{2} - 1\right]$, i.e., during the first two weeks of the month: If $x = 1$, 2 consecutive days off to j from h. If $x = 2$, 2 non-consecutive days off to j from h. When $d \in \left[\frac{numDays}{2}, numDays - 1\right]$, i.e., during the last two weeks of the month: If $x = 1$, 2 non-consecutive days off to j from h. If $x = 2$, 2 consecutive days off to j from h.

4.4 Remaining Days Within Each Period

The procedure for the remaining days within each period is described in Algorithm 2. This algorithm represents the behaviour of the heuristic during the first two weeks and the last two weeks of the month, excluding the first day of each period. The input parameters include the daily staffing requirements and the lists of available, busy, resting, and off-duty employees. In addition, the list of shifts to be assigned is created during the execution of the algorithm.

Similarly to the procedure used on the first day of the first period, the staffing requirements are scanned for each shift starting time to determine whether additional employees must be assigned. The heuristic first attempts to assign employees to the same shift as on the previous working day, using $ReturnPreviousShift$, in order to promote schedule stability. This process is repeated until either all requirements are satisfied or no suitable employees remain available.

If the staffing requirements are not fully satisfied, the corresponding shift starting times are recorded in $ShiftsToAssign$, with repetition whenever more

Algorithm 1 First day of the first period

Input: $L_{AVAIL}, L_{BUSY}, L_{REST}, MinReq_{h,d}$
for all $h \in [0, 48[$, with $d = 0$ **do**
  $numWorking \leftarrow \text{SIZE}(L_{BUSY})$
  if $numWorking < MinReq_{h,d}$ **then**
    $numRequired \leftarrow MinReq_{h,d} - numWorking$
    for $w \in [0, numRequired[$ **do**
      $e \leftarrow \text{SELECTEMPLOYEE}(L_{AVAIL})$
      $(S_F, P_F) \leftarrow \text{ASSIGNSHIFT}(e, h)$
      $L_{BUSY} \leftarrow L_{BUSY} \cup \{e\}$
      $L_{AVAIL} \leftarrow L_{AVAIL} \setminus \{e\}$
  for all $b \in L_{BUSY}$ **do**
    $shiftEnd \leftarrow \text{RETURNSHIFTEND}(b)$
    if $h = shiftEnd$ **then**
      $L_{REST} \leftarrow L_{REST} \cup \{b\}$
      $L_{BUSY} \leftarrow L_{BUSY} \setminus \{b\}$
  for all $r \in L_{REST}$ **do**
    $restEnd \leftarrow \text{RETURNRESTEND}(r)$
    if $h = restEnd$ **then**
      $L_{AVAIL} \leftarrow L_{AVAIL} \cup \{r\}$
      $L_{REST} \leftarrow L_{REST} \setminus \{r\}$
Output: $(S, P) = \bigcup(S_F, P_F)$

than one additional employee is required. The remaining employees in L_{AVAIL} are then assigned, in their current order, to the shift starting times stored in that list. In this process, a counter k is used to track the number of assigned employees for each shift, ensuring that the required number of employees is reached.

As in the first day, the lists of busy and resting employees are continuously updated to reflect the completion of shifts and the end of rest periods. Days off are assigned based on each employee's work history. Employees who have worked five consecutive days are identified through $History$, and the function $AssignDaysOff$ is then applied according to the employee's identification number and the current period of the month. As defined in Table 1, the interpretation of parameter x differs between the first and second halves of the month, ensuring a balanced distribution of consecutive and non-consecutive days off.

Finally, the list of employees on a day off is updated so that employees who complete their days off are returned to the list of available employees.

4.5 First Day of the Second Period

The procedure for the first day of the second period is similar to that used on the first day of the first period, since the objective is to promote schedule stability throughout the month, *i.e.*, to avoid large variations in the shifts assigned to each employee over periods of up to two weeks. However, at this stage, the list of available employees no longer contains all employees of a given team. Therefore,

Algorithm 2 Remaining days within each period

Input: $L_{AVAIL}, L_{BUSY}, L_{REST}, L_{OFF}, MinReq_{h,d}, numDays$

for $d = 1$ to $numDays - 1$ **do**
  if $d \neq \frac{numDays}{2}$ **then**
    for $h = 0$ to 48 **do**
      $numWorking \leftarrow$ Size(L_{BUSY})
      if $numWorking < MinReq_{h,d}$ **then**
        $numRequired \leftarrow MinReq_{h,d} - numWorking$
        $k \leftarrow 0$
        $ShiftsToAssign \leftarrow \emptyset$
        for all $v \in L_{AVAIL}$ **do**
          $prevShift \leftarrow$ ReturnPreviousShift(v)
          if $k < numRequired$ **then**
            if $h = prevShift$ **then**
              $(S_F, P_F) \leftarrow$ AssignShift(v, h)
              $L_{BUSY} \leftarrow L_{BUSY} \cup \{v\}$
              $L_{AVAIL} \leftarrow L_{AVAIL} \setminus \{v\}$
              $k \leftarrow k + 1$
        while $k < numRequired$ **do**
          $ShiftsToAssign \leftarrow ShiftsToAssign \cup \{h\}$
          $k \leftarrow k + 1$
        for all $q \in ShiftsToAssign$ **do**
          if Size(L_{AVAIL}) > 0 **then**
            $v \leftarrow$ SelectEmployee(L_{AVAIL})
            $(S_F, P_F) \leftarrow$ AssignShift(v, q)
            $L_{BUSY} \leftarrow L_{BUSY} \cup \{v\}$
            $L_{AVAIL} \leftarrow L_{AVAIL} \setminus \{v\}$
      for all $l \in L_{BUSY}$ **do**
        $shiftEnd \leftarrow$ ReturnShiftEnd(l)
        if $h = shiftEnd$ **then**
          $L_{REST} \leftarrow L_{REST} \cup \{l\}$
          $L_{BUSY} \leftarrow L_{BUSY} \setminus \{l\}$
      for all $g \in L_{REST}$ **do**
        $restEnd \leftarrow$ ReturnRestEnd(g)
        if $h = restEnd$ **then**
          $hist \leftarrow$ History(g)
          $histSize \leftarrow$ Size($hist$)
          if $histSize = 5$ **then**
            $empId \leftarrow$ ReturnId(g)
            if $empId \bmod 2 = 0$ **then**
              $(S_F, P_F) \leftarrow$ AssignDaysOff($g, h, d, 1$)
              $L_{OFF} \leftarrow L_{OFF} \cup \{g\}$
              $L_{REST} \leftarrow L_{REST} \setminus \{g\}$
            else
              $(S_F, P_F) \leftarrow$ AssignDaysOff($g, h, d, 2$)
              $L_{OFF} \leftarrow L_{OFF} \cup \{g\}$
              $L_{REST} \leftarrow L_{REST} \setminus \{g\}$
          else
            $L_{AVAIL} \leftarrow L_{AVAIL} \cup \{g\}$
            $L_{REST} \leftarrow L_{REST} \setminus \{g\}$
      for all $f \in L_{OFF}$ **do**
        $offEnd \leftarrow$ ReturnDayOffEnd(f)
        if $(h, d) = offEnd$ **then**
          $L_{AVAIL} \leftarrow L_{AVAIL} \cup \{f\}$
          $L_{OFF} \leftarrow L_{OFF} \setminus \{f\}$

Output: $(S, P) = \bigcup(S_F, P_F)$

reordering only this list would have a limited effect on the variability of the resulting assignments. For this reason, all employee lists are randomly reordered before the procedure is applied.

In addition, days off are taken into account in this procedure. The mechanism for assigning days off follows the same rule used for the remaining days of the second period: employees with an odd identification number are assigned two consecutive days off, whereas employees with an even identification number are assigned two non-consecutive days off.

5 Local Search Strategies

5.1 Evaluation Functions

For this case study, two objective functions were defined. The objective function $z_1(S, P)$ is defined as the percentage of times employees maintain the same shift over each two-week period, i.e., within each of the two periods of the month. Thus, $z_1(S, P)$ is computed as the average of the ratios between the number of identical shift starting times assigned to each employee in each period and the total number of shifts assigned to that employee in the corresponding period. More specifically, the value of $z_1(S, P)$ is computed as follows:

1. For each employee, compute the number of identical shift starting times in the first period of the month;
2. For each employee, divide this value by the total number of assigned shifts in that period, obtaining a relative frequency;
3. Compute the average of these values over all employees;
4. Repeat steps 1–3 for the second period of the month;
5. Compute the average of the two values obtained.

Therefore, $z_1(S, P)$ is defined in Equation (1) where $t_{i,1}$ and $t_{i,2}$ denote the number of identical shift starting times assigned to employee i in periods 1 and 2, respectively, and $Total_{i,1}$ and $Total_{i,2}$ denote the total number of shifts assigned to employee i in those periods. A solution is considered better when the average percentage of identical shifts assigned to each employee across both periods is higher.

$$z_1(S, P) = \frac{1}{2} \cdot \frac{\sum_{i=1}^{|S|} \left(\frac{t_{i,1}}{Total_{i,1}} + \frac{t_{i,2}}{Total_{i,2}} \right)}{|S|} \cdot 100 \tag{1}$$

After obtaining initial solutions for several instances, it was observed that the workload distribution among employees could be significantly unbalanced. In order to promote fairness and equity, a second objective function, $z_2(S, P)$, was defined, corresponding to the mean absolute deviation of the number of non-working days assigned to each employee. That is, it measures the average absolute deviation between the number of non-working days assigned to each employee and the average number of non-working days across all employees in the team. In Equation (2), $z_2(S, P)$ is defined, where x_i denotes the number of non-working days assigned to employee i and $\bar{x}$ denotes the average number of

non-working days across the team. A solution (S_A, P_A) is considered better than a solution (S_B, P_B) if $z_2(S_A, P_A) < z_2(S_B, P_B)$.

$$z_2(S, P) = \frac{\sum_{i=1}^{|S|} |x_i - \bar{x}|}{|S|} \tag{2}$$

5.2 Neighborhood Structures

Two neighbourhood structures were defined for generating neighbouring solutions (S', P') from a solution (S, P), corresponding to two distinct local search methods.

$\mathbf{NS}_1$: Shift swap between two employees. The neighbouring solutions of a solution (S, P) in neighbourhood $\mathbf{NS}_1(S, P)$ consist of all feasible solutions obtained by swapping one shift between two employees on a given day or set of days within a month.

$\mathbf{NS}_2$: Swap of one unassigned day and one shift between two employees. The neighbouring solutions of a solution (S, P) in neighbourhood $\mathbf{NS}_2(S, P)$ consist of all feasible solutions obtained by swapping one shift and one day with no assignment between two employees on a given day or set of days within a month. It should be noted that a day is considered to have no assignment for a given employee if that employee is neither assigned to any shift nor allocated a day off on that day.

The main differences between the two neighbourhood structures lie in the conditions that must be satisfied in each employee's schedule on a given day in order for exchanges, and consequently changes, to be performed, as well as in the criterion used to select the best solution, namely the corresponding evaluation function (z_1 for the first neighbourhood structure and z_2 for the second one).

5.3 Local Search Algorithms

Two local search algorithms were implemented, one for each objective function and neighbourhood structure. In both cases, the procedure terminates when no improving solution exists in the neighbourhood. In these approaches, the next current solution is selected as the neighbouring solution with the best value of the corresponding evaluation function.

Local Search Method 1 (LS1). Starting from the initial solution (S, P), algorithm LS1 searches for neighbouring solutions using neighbourhood structure $\mathbf{NS}_1$. Initially, the shift assigned to employee x_1 on the first day is swapped with the shift assigned to employee x_2 on that day, and the value of the evaluation function z_1 is computed for the resulting solution. This procedure is repeated for the remaining employees, i.e., the shift of employee x_1 on day 1 is swapped with the shifts of all other employees assigned to a shift on that day, and the value of z_1 is computed for each neighbouring solution obtained. The best solution is then selected and becomes the current solution.

After the new current solution is selected, the procedure described above is repeated for the shift of employee x_2 on the first day of the month and, in

the same way, for all employees assigned to a shift on that day. Once the best solution for that day has been identified, it is no longer modified, i.e., the shifts assigned to each employee on day 1 remain fixed until the end of the local search. The procedure is repeated for all employees assigned to a shift and for all days of the month for which such exchanges are possible, until the stopping criterion is satisfied. The solution obtained is therefore a local optimum.

Local Search Method 2 (LS2). Starting from the initial solution (S, P), algorithm LS2 searches for improved solutions through neighbourhood structure $\mathbf{NS}_2$. Initially, an exchange is performed between the schedule of employee x_1 on day 1 and the schedules of employees $x_2, x_3, x_4, \ldots, x_n$ on that same day, and, in this case, the value of the evaluation function z_2 is computed for each neighbouring solution. The best solution is then selected and becomes the current solution. This procedure is repeated for the schedule of employee x_2 on day 1, for all remaining employees, and for all other days of the month under analysis.

However, since the aim is to reduce workload imbalance among employees, i.e., the deviation between the number of non-working days assigned to each employee in that month and the average number of non-working days of the employees in the corresponding team, it is not relevant to exchange the schedules of two employees who, on the day under consideration, are both assigned to some shift, both have a non-working day, or one of them has a day off assigned. Consequently, exchanges are only performed between two employees when one of them is assigned to a specific shift on the day under consideration and the other is associated with a day with no assignment. As mentioned above, this procedure is repeated for all days of the month for which such exchanges are possible, until the stopping criterion is satisfied. The solution obtained is therefore a local optimum.

6 Computational Experiments

6.1 Instances and Computational Setting

The proposed approaches were tested using 39 instances. Ten of these instances are based on real data from a real company, although some data were modified for confidentiality reasons. The remaining 29 instances were randomly generated, based on the instances used in [11]. These instances consider different employee teams and the corresponding staffing requirements for each day across different months of the year. Each instance consists of a spreadsheet specifying the minimum number of employees required for each 30-minute period – corresponding to the starting time of each of the 48 shifts – for every day of the month under consideration. Additionally, the spreadsheet includes information about the team to be scheduled, such as employee names, unique identification numbers, and skill sets. All structures and algorithms were implemented in the C# programming language and executed on a computer equipped with an Intel Core i7-8550U processor at 1.80 GHz and 16 GB of RAM.

6.2 Results and Discussion

Table 2 presents the computational results obtained. All instances (column i) are evaluated according to the same criteria: the number of employees in each instance (n), the number of days in the planning horizon (d), the execution time in seconds (t_{exc}), the values of the objective functions (z_1 and z_2), the percentage improvement obtained for z_1 using algorithm LS1 (imp_{LS1}), and the percentage improvement obtained for z_2 using algorithm LS2 (imp_{LS2}).

From the analysis of the results obtained, it can be concluded that, for the initial solutions, the average value of z_1 is approximately 71.1%. This means that, across the different instances, employees maintain the same shift, on average, in 71.1% of the cases within each period, i.e., every two weeks. For these same instances, considering the solutions generated by the CSI constructive heuristic, the average value of z_2 is approximately 1.7, which corresponds to the mean absolute deviation in the number of non-working days assigned to employees in a given month.

Still considering the initial solutions, the instance with the highest value of z_1 (98.9%) is composed of 46 employees and corresponds to a 31-day month. For this instance, the value of z_2 is 2.7, which is substantially lower than the average. The instance with the lowest value of z_2 (0.7) is also composed of 46 employees and corresponds to a 28-day month.

The results obtained with local search method 1 (LS1) show that the value of z_1 for the initial solution is improved in 33 out of the 39 test instances and remains unchanged in the remaining instances, whereas the value of z_2 remains unchanged for all instances. The 6 instances in which the value of z_1 remains unchanged are those in which there is excess staffing, i.e., the number of employees is too high relative to the organisation's requirements. Since the number of employees is much higher than the staffing requirements, the CSI heuristic is able to produce schedules in which employees maintain the same shifts a high number of times every two weeks and, therefore, the resulting solution tends to be of good quality. In contrast, LS1 is unable to further improve the solution, since shifts are already relatively well distributed in the initial solution. The fact that the value of the objective function z_2, under LS1, remains equal to that of the initial solution in all instances can be explained by the way the neighbourhood structures were constructed. Algorithm LS1 operates on the first neighbourhood structure and aims to maximise the number of identical shifts assigned to each employee every two weeks. Therefore, in this algorithm, exchanges are only performed between two employees if both are assigned to some shift on the day under consideration. On the other hand, objective function z_2 computes the sum of the deviations in the number of non-working days assigned to each employee. Since exchanges are only performed between employees who are both assigned to shifts, the value of z_2 remains unchanged, because no employee moves from being assigned to being unassigned or associated with a day off. Employees only exchange one shift for another. Thus, LS1 cannot affect the value of z_2, which is confirmed by the reported results.

Table 2. Computational results for CSI, LS1 and LS2

			CSI			LS1				LS2			
i	n	d	z_1	z_2	t_{exc}	z_1	z_2	t_{exc}	imp_{LS1}	z_1	z_2	t_{exc}	imp_{LS2}
1	46	28	74.0	0.70	0.71	80.0	0.70	4.90	8.1	73.9	0.35	1.90	50.0
2	46	28	76.0	0.87	0.66	83.3	0.87	4.40	9.6	75.7	0.57	1.43	34.5
3	25	28	70.8	1.00	0.50	76.2	1.00	3.79	7.6	70.3	0.68	0.27	32.0
4	9	28	69.5	2.22	1.45	72.9	2.22	0.40	4.9	69.2	0.89	0.27	59.9
5	17	28	74.9	0.76	0.35	80.3	0.76	1.30	7.2	75.5	0.53	0.32	30.3
6	8	28	72.6	1.38	0.27	79.8	1.38	0.36	9.9	73.3	0.63	0.29	54.3
7	10	28	75.6	1.60	0.27	78.9	1.60	0.37	4.4	75.0	1.00	0.30	37.5
8	11	28	61.4	1.09	0.37	67.2	1.09	0.61	9.4	61.2	0.91	0.29	16.5
9	24	31	70.6	1.21	0.30	76.6	1.21	2.78	8.5	70.4	0.88	0.53	27.3
10	46	31	75.8	1.24	0.30	79.7	1.24	19.52	5.0	74.9	0.76	1.42	38.7
11	19	31	58.8	1.74	0.20	67.3	1.74	2.45	14.5	58.7	1.53	0.47	12.1
12	46	31	94.6	3.72	0.39	94.6	3.72	1.01	0.0	78.7	0.63	3.03	83.1
13	46	31	89.1	3.61	0.33	89.1	3.61	1.20	0.0	77.2	0.57	3.88	84.2
14	46	31	92.4	3.33	0.29	92.4	3.33	1.32	0.0	79.9	0.59	3.20	82.3
15	46	31	98.9	2.70	0.32	98.9	2.70	1.50	0.0	90.5	0.65	3.91	75.9
16	46	31	92.4	3.52	0.33	92.4	3.52	1.28	0.0	78.7	0.61	4.23	82.7
17	25	31	87.0	2.20	0.39	90.0	2.20	1.25	3.4	86.3	1.88	3.35	14.5
18	25	31	82.4	1.92	0.44	85.4	1.92	1.39	3.6	82.0	1.68	3.45	12.5
19	25	31	78.5	1.64	0.45	81.6	1.64	1.18	3.9	77.0	1.00	3.46	39.0
20	25	31	74.0	1.44	0.32	79.1	1.44	1.18	6.9	73.7	1.04	2.94	27.8
21	25	31	76.1	1.52	0.32	80.4	1.52	1.09	5.7	76.1	1.44	3.57	5.3
22	24	31	82.1	2.13	0.30	85.3	2.13	1.11	3.9	82.1	2.04	1.14	4.2
23	24	31	79.8	1.92	0.37	82.7	1.92	1.10	3.6	77.6	1.33	0.74	30.7
24	24	31	71.5	1.71	0.28	77.8	1.71	1.21	8.8	69.0	0.71	0.67	58.5
25	24	31	71.6	1.42	0.44	75.6	1.42	1.52	5.6	70.7	0.83	0.96	41.5
26	24	31	74.1	1.75	0.82	79.2	1.75	1.60	6.9	71.5	0.75	1.20	57.1
27	17	31	49.5	1.47	0.37	57.8	1.47	1.34	16.8	49.5	1.35	0.87	8.2
28	17	31	42.8	0.71	0.37	53.8	0.71	1.15	25.7	42.7	0.59	0.80	16.9
29	17	31	37.7	1.12	0.22	50.6	1.12	0.77	34.2	37.8	0.76	0.51	32.1
30	17	31	43.6	1.29	0.20	54.0	1.29	0.72	23.8	43.5	0.94	0.47	27.1
31	17	31	48.8	1.18	0.22	60.9	1.18	0.74	24.8	48.8	1.06	0.46	10.2
32	46	31	95.7	2.78	0.45	95.7	2.78	0.65	0.0	88.8	1.26	0.49	54.7
33	25	28	94.6	1.44	0.24	95.8	1.44	1.65	1.3	91.9	0.16	0.53	88.9
34	25	31	72.9	0.92	0.33	76.1	0.92	0.74	4.4	72.7	0.76	0.46	17.4
35	19	31	57.5	1.63	0.71	66.6	1.63	1.30	15.8	55.8	0.68	1.10	58.3
36	19	31	50.2	1.47	0.42	59.3	1.47	1.43	18.1	50.0	1.16	0.91	21.1
37	19	31	48.9	1.32	0.39	59.9	1.32	1.17	22.5	48.3	0.68	0.92	48.5
38	19	31	52.1	0.95	0.31	54.5	0.95	0.82	4.6	51.4	0.63	0.81	33.7
39	19	31	54.8	1.53	0.29	60.7	1.53	0.74	10.8	54.8	1.42	0.49	7.2
Avg			71.1	1.7	0.40	76.2	1.7	1.9	8.8	68.4	0.9	1.4	38.9

Considering the application of algorithm LS2, the value of z_1 for the initial solution improves in only 3 of the 39 test instances and remains unchanged in 5 instances, while in the remaining cases the value of z_1 decreases. In contrast, the value of z_2 improves in all test instances. These results can also be explained by the way LS2 operates. Algorithm LS2 uses the second neighbourhood structure, which was designed to minimise the mean deviation in the number of non-working days assigned to each employee, either because the employee is not assigned to any shift or because the employee is associated with a day off. In this algorithm, exchanges are only performed between two employees when one of them is assigned to a shift on the day under consideration and the other is neither assigned to a shift nor associated with a day off on that same day. Hence, the value of z_1 is unlikely to improve under LS2, since shifts that contribute positively to increasing the value of z_1 may be exchanged. As a consequence, the value of objective function z_1 may even deteriorate.

The results obtained for all instances with respect to z_1 show an average improvement of 8.8% using LS1, with an average computational time of approximately 0.4 seconds. The smallest average improvement is observed for an instance with a team of 25 employees and a 28-day month (1.3%), with a computational time of 1.65 seconds. The largest average improvement is observed for an instance corresponding to the scheduling of 17 employees over a 31-day month (34.2%), obtained in approximately 0.77 seconds.

The computational results obtained with LS2 with respect to z_2 reveal an average improvement of 38.9%, with an average computational time of approximately 1.4 seconds. In this algorithm, the smallest average improvement is observed for an instance composed of 24 employees and a 31-day month (4.2%), with a computational time of 1.14 seconds. In turn, the largest average improvement is observed for an instance composed of 25 employees and a 28-day month (88.9%), obtained in 0.53 seconds. By analysing the instances in which the improvements in the value of z_2 are greatest, it can be seen that these are the same instances for which the value of z_1 remains unchanged under LS1. As noted above, in these instances the number of employees appears to be too high relative to the organisation's requirements and, consequently, several employees remain without any assignment because they are not required to meet demand. Thus, the initial value of z_2 is relatively high. However, after applying LS2, the value of z_2 is substantially reduced, since the algorithm operates on the second neighbourhood structure.

It is also worth noting that the smallest average improvement in z_1 obtained with LS1 and the largest average improvement in z_2 obtained with LS2 are observed for the same instance. More precisely, the smallest average improvement in z_1 under LS1 is obtained for instance 33, which is also the instance for which the largest average improvement in z_2 is achieved under LS2.

The computational results also reveal a trade-off between shift stability and workload fairness. In particular, LS1 improves the stability-related objective function z_1 without affecting z_2, whereas LS2 significantly improves the fairness-related objective function z_2 but may deteriorate z_1. This behaviour reflects the

conflicting nature of the two objectives: increasing schedule stability tends to preserve existing assignments, while improving workload balance often requires modifying them.

7 Conclusion and Future Work

This work addresses a personnel scheduling problem arising in a call center service operating 24 h a day, 7 d a week. In the call center under analysis, the scheduling process is currently performed manually. Therefore, the main objective of this paper was the development and implementation of a heuristic method capable of constructing high-quality solutions while reducing the time required for this process. To this end, heuristic approaches for assigning shifts to employees were proposed. First, a constructive heuristic was implemented to obtain initial solutions. In addition, two local search methods were implemented with the aim of improving these initial solutions. The constructive heuristic and the two local search methods were tested on real instances, modified for confidentiality reasons, as well as on randomly generated instances. The constructive heuristic produced solutions of relatively good quality, i.e., with satisfactory values for the two objective functions in the context of the problem. The two local search methods yielded significant improvements in the great majority of cases.

As future work, it would be interesting to adopt a multiobjective strategy, i.e., to combine the two objective functions into a single one by assigning weights to each criterion. Moreover, it would also be important to take employee preferences regarding shift and day-off assignment into account and, in this context, to analyse other performance measures, such as employee satisfaction.

Acknowledgments. This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Unit Project Scope UID/00319/2025 - Centro ALGORITMI (ALGORITMI/UM) https://doi.org/10.54499/UID/00319/2025

References

1. Atlason, J., Epelman, M.A., Henderson, S.G.: Call center staffing with simulation and cutting plane methods. Ann. Oper. Res. **127**(1), 333–358 (2004)
2. Avramidis, A.N., Chan, W., Gendreau, M., L'ecuyer, P., Pisacane, O.: Optimizing daily agent scheduling in a multiskill call center. Eur. J. Oper. Res. **200**(3), 822–832 (2010)
3. Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., De Boeck, L.: Personnel scheduling: a literature review. Eur. J. Oper. Res. **226**(3), 367–385 (2013)
4. Brucker, P., Qu, R., Burke, E., Post, G.: A decomposition, construction and post-processing approach for nurse rostering. In: 2nd Multidisciplinary International Conference on Scheduling: Theory & Applications, MISTA 2005, pp. 397–406. University of New York (2005)
5. Burke, E.K., De Causmaecker, P., Berghe, G.V., Van Landeghem, H.: The state of the art of nurse rostering. J. Sched. **7**(6), 441–499 (2004)

6. Cezik, M.T., L'Ecuyer, P.: Staffing multiskill call centers via linear programming and simulation. Manage. Sci. **54**(2), 310–323 (2008)
7. Chevalier, P., Van den Schrieck, J.C.: Optimizing the staffing and routing of small-size hierarchical call centers. Prod. Oper. Manag. **17**(3), 306–319 (2008)
8. Fukunaga, A., Hamilton, E., Fama, J., Andre, D., Matan, O., Nourbakhsh, I.: Staff scheduling for inbound call and customer contact centers. AI Mag. **23**(4), 30–30 (2002)
9. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: tutorial, review, and research prospects. Manuf. Serv. Oper. Manage. **5**(2), 79–141 (2003)
10. Hast, J.: Optimal Work Shift Scheduling: A Heuristic Approach. Master's thesis, Aalto University, School of Science (2017)
11. Martins, R., Pinto, T., Alves, C.: An exact optimization approach for personnel scheduling problems in the call center industry. In: International Conference on Computational Science and Its Applications, pp. 407–424. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37108-0_26
12. Pot, A., Bhulai, S., Koole, G.: A simple staffing method for multiskill call centers. Manuf. Serv. Oper. Manage. **10**(3), 421–428 (2008)
13. Robbins, T.R., Harrison, T.P.: A stochastic programming model for scheduling call centers with global service level agreements. Eur. J. Oper. Res. **207**(3), 1608–1619 (2010)
14. Silva, A.V.: Algoritmos heurísticos para escalonamento de pessoal em call centers [Heuristic algorithms for personnel scheduling in call centers]. Master's thesis, University of Minho (2021), in Portuguese
15. Türker, T., Demiriz, A.: An integrated approach for shift scheduling and rostering problems with break times for inbound call centers. Math. Probl. Eng. **2018**(1), 7870849 (2018)

Distributionally Robust Optimisation for Prepositioning Wildfire Fighting Resources

Agostinho Agra[1], Filipe Alvelos[2(✉)], Francisco Marques[3], Ana Raquel Xambre[4], and Helena Alvelos[4]

[1] Departamento de Ciências Matemáticas/CEMS.UL, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
amagra@fc.ul.pt

[2] Department of Production and Systems/ALGORITMI Research Center/LASI, School of Engineering, University of Minho, Guimarães, Portugal
falvelos@dps.uminho.pt

[3] Department of Mathematics, Department of Economics, Management, Industrial Engineering and Tourism, University of Aveiro, Aveiro, Portugal
franciscocmarques@ua.pt

[4] Department of Economics, Management, Industrial Engineering and Tourism/Centre for Research & Development in Mathematics and Applications, University of Aveiro, Aveiro, Portugal
{raquelx,helena.alvelos}@ua.pt

Abstract. We address the problem of prepositioning firefighting resources for wildfire suppression. Given the inherent uncertainty in the location of ignitions and the wind, we formulate a two-stage optimisation model. In the first stage, resource prepositioning decisions are made in anticipation of possible fire outbreaks; in the second stage, once ignitions are realised, the movement of resources to attack positions is determined. Fire spread is modeled using the minimum travel time principle and is represented through a mixed-integer programming model that integrates both fire spread and operational decisions.

To address uncertainty, we propose a two-stage distributionally robust optimisation model which includes a discrete set of scenarios characterised by varying wind directions and ignition locations. A row-and-column decomposition algorithm is developed to efficiently solve the resulting large-scale problem.

Computational experiments are conducted on realistic instances derived from an actual landscape. Rothermel's fire behaviour model is used to estimate the fire potential transmission times. Each instance considers 80 scenarios reflecting different ignition patterns and wind behaviours. The results demonstrate the effectiveness of the proposed approach and highlight the scalability and computational efficiency of the decomposition method.

Keywords: Fire suppression · Distributionally robust optimisation · Two-stages · Decomposition algorithm

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 283–300, 2027.
https://doi.org/10.1007/978-3-032-30521-3_18

1 Introduction

Wildfires are natural disasters that can have a significant impact on ecosystems which can lead to the loss of lives and material damage. Therefore, their management is a highly relevant issue. Following [1], wildfire management can be divided in three main areas: prevention, detection and management of fire suppression resources. In this paper, we address an operational management of fire suppression resources problem. Although the first operational research approaches to this type of problems are from the early 1960 s [2], only recently a categorized literature review [3] and a survey [4] have been published, supporting the current relevance of the topic.

Operational management of fire suppression resources can be typified in deployment, dispatch and positioning. Deployment decisions are taken before an actual fire is detected and typically amount to defining which resources should be (re)allocated to which bases. Dispatch decisions are taken after an ignition is detected and amount to deciding which resources should be sent to the fire. Lastly, positioning decisions detail where the dispatched resources should attack the fire.

In this paper, we consider a problem that combines the prepositioning and positioning of resources. The decisions of repositioning the resources, which can be seen as deployment decisions but with allocation alternatives other than the bases - for example, roads - are taken before fire is detected. After a fire is detected, the attack positions of the resources are defined. Given the uncertainty involved in the location of the fire (and also on the wind), we derive an optimisation under uncertainty approach based on mixed integer programming (MIP).

Different MIP models have been proposed for the positioning problem. In [5], the minimum travel time principle (MTT, [6]) was used for the first time to model the fire paths and arrival times in a landscape. According to this principle, fire follows the quickest path between any two locations of a landscape. This principle allows to model fire spread as shortest paths in a network and has since been used in software that simulates fire spread (e.g. [7]) and in MIP models for different variants of the positioning problem, for example in [3,8–12].

These are deterministic approaches that, at the expense of increasing the complexity and difficulty of the approach, can be extended to address uncertainty.

Stochastic programming (SP) has been applied to problems related to the one addressed in this paper with Distributionally Robust Optimisation (DRO). For example, in [13], a two-stage model is proposed to minimise the total cost (that includes the cost of the used resources, the expected suppression cost and the net value change). Scenarios are characterised by a set of fires, each with a known perimeter at each time period, and, among other attributes, by the line production rate of the resources. Ntaimo et al. [14] also proposed a two-stage SP model with a similar objective function. A major difference is that the effectiveness of the suppression is assessed by the number of fires that do not receive a standard response, instead of fire containment measures. In [15], a two-

stage MIP model formulates the problem of defining the bases for resources given a set of scenarios, each characterised by the number fires that occur on a day, along with their location and their ignition time. For each scenario, a standard response, i.e. the number of resources required to extinguish the fires is derived. The model minimises the expected value (each scenario has a probability) of fires that do not receive a standard response. A detailed case study with data from California, with a rigorous parametrisation of the model (including the generation of scenarios and resources) is discussed. SP has also been applied in other Wildfire problems, where for defining attacking positions with wind uncertainty [16,17].

A major difference of the proposed approach for prepositioning, besides the use of DRO instead of SP, lies in how fire spread is explicitly modelled. In the proposed approach, fire spread is integrated with decisions related to resource prepositioning and attack positioning. This type of *integrated* approach has also been used in related problems with Robust Optimisation (RO). Mendes and Alvelos [18] proposed a robust optimisation approach to address uncertainty on the number of resources required to suppress fire. Marques et al. [19] addressed the same problem as the one of this work with "pure" robust optimisation.

DRO may be seen as a generalisation that encompasses SP and RO as special cases. While SP assumes that the true probability distribution of the uncertain parameters is known, RO only assumes that the parameters belong to an uncertainty set. In addition, while SP optimises over the expected value, RO evaluates the decisions under the worst-case scenario. In DRO, it is assumed that the probability distribution is unknown but belongs to a set called the ambiguity set. In our case, we use an extension of the one-norm Wasserstein ambiguity set, also known as the Kantorovich ambiguity set. This set is centred on the nominal distribution and includes all distributions whose distance from the nominal distribution does not exceed a parameter ϵ. To measure the distance, we use a transportation problem (where the quantities associated with the origins correspond to the probability mass of given scenarios in the nominal distribution and the quantities in the destination correspond to the probability mass of the final distribution), subject to an additional knapsack constraint that controls the mass to transport. The goal is to determine the solution that minimises the expected cost considering the worst-case probability distribution within the ambiguity set. When the ambiguity set uses the Kantorovich ambiguity set and the support is discrete, Bansal et al. [20] show that two-stage DRO mixed-integer problems can be converted into (larger) computationally tractable single-stage mixed-integer problems. For reviews on DRO, see [21,22].

Given its flexibility, DRO is becoming very popular. To the best of our knowledge, it has never been applied to wildfire management. For a related application with wind uncertainty, see [23].

In this paper, we propose a two-stage DRO model based on a discrete set of scenarios. These scenarios are defined by a wind direction and the locations of one or more ignitions. Both factors result from two statistical studies conducted on each factor alone [24]. We consider that the spatial distribution of fuels and

slopes of the landscape under study are known. The problem is to decide where to prepositioning the resources such that the burned area is minimized in the worst-case scenario. The approach can be seen as having static variables (related to the prepositioning decisions) and adjustable variables (related to the decisions of each scenario). Both the number of constraints and the number of adjustable variables depend on the number of scenarios. To solve the problem we propose a row-and-column decomposition algorithm. For a comparison of the different solution approaches for the two-stage DRO problem under different ambiguity sets, see [25].

This paper is organized as follows: we provide a description of the problem in Sect. 2. In Sect. 3, we introduce the DRO model. The row-and-column decomposition algorithm is presented in Sect. 4. The computational experiments conducted to provide insights on the solutions and to evaluate the proposed model and decomposition algorithm are reported in Sect. 5. Finally, in Sect. 6, we provide the conclusions and future research directions.

2 Problem Definition, Notation and a Deterministic Model

We consider the problem of prepositioning a set of available firefighting resources, possibly of different types, in a landscape before a wildfire ignition occurs with the objective of minimising the worst-case burned area, within a given time period, of a set of scenarios, each one characterised by one or more ignitions and a wind speed and direction. After the prepositioning decision is taken and the uncertainty is disclosed (i.e. the actual scenario is known), the resources can be moved from their prepositioning locations to the ones where they interdict fire spread.

In the context of optimisation under uncertainty, the problem considered is a two-stage problem. In the first-stage, the prepositioning of the resources is decided, i.e. static decisions are taken. Then the uncertainty (ignition(s) locations and wind) is revealed. In the second stage, the (adjustable) decisions regarding movement from prepositioning positions to the positions where the resources attack fire are made.

To model this problem, we represent the landscape as a network in which nodes correspond to centroids of hexagons forming a grid, and arcs represent potential fire transmission between adjacent nodes. Each arc has a fire transmission time that corresponds to the time it takes for the fire to reach the destination node of the arc once the fire reached the origin node. In our computational experiments, the fire transmission times of the arcs were derived from the fire behaviour model of Rothermel [26], which accounts for fuels, slopes, and winds to estimate the fire rate of spread under homogeneous conditions. It is worth noting that, in the optimisation model, the fire transmission time is a parameter and therefore can be obtained with other methods.

Based on the fire transmission times, it is possible to model the fire spread by using the minimum travel time principle [6], which states that the time it takes

for the fire to reach a node is the length (duration) of the shortest (quickest) path from an ignition node.

We now introduce the notation used.

The fire network, already introduced, is represented by $G^f(N^f, A^f)$. The set of nodes that are burning at the detection time, e.g. the fire perimeter, is represented by I^f. The fire transmission time between adjacent nodes i and j (i.e. of the arc ij) is represented by c_{ij}. The instant where the burned area is evaluated is denoted by T.

We consider a set of available resource types defined by R. Each resource type, r, $r \in R$, is associated with a graph, represented by $G^r(N^r, A^r)$, where the nodes correspond to potential attack positions and the arcs to potential movements.

The fire and resource nodes are related through a set of pairs (l, i), denoted by A^{rf}, where l is a node of the resource network and i is a node of the fire node. If a unit of resource r is positioned at a node l of its resource network, then all the nodes i of the fire network for which $(l, i) \in A^{rf}$ are attacked. A fire node being attacked means that fire will no longer spread through the outgoing arcs of i. The constant d^r_{kl} is the quickest path duration of resource r between nodes k and l. The number of available resource units of type r is denoted by b^r.

We now introduce a deterministic model that assumes that the prepositioning of the resources is known and provided by parameters $\bar{x}^r_k = 1$ if a resource of type r is prepositioned at node k and $\bar{x}^r_k = 0$, otherwise. If no uncertainty is present, this model provides the optimal attack positions for all resources, given their current location (i.e. assuming fixed prepositions). The following decision variables are defined:

$$t_i \text{ fire arrival time at node } i \qquad i \in N^f;$$

$$z^r_{kl} = \begin{cases} 1, & \text{if the resource of type } r \text{ prepositioned at node } k \text{ is moved to node } l \\ 0, & \text{otherwise} \end{cases} \qquad kl \in A^r, r \in R;$$

$$y_i = \begin{cases} 1, & \text{if node } i \text{ burns before instant } T \\ 0, & \text{otherwise} \end{cases} \qquad i \in N^f.$$

The MIP deterministic model is:

$$\text{Min} \sum_{i \in N^f} y_i \tag{1}$$

s.t.

$$t_i = 0 \qquad i \in I^f \tag{2}$$

$$t_j \leq t_i + c_{ij} + T \sum_{kl \in A^r, li \in A^{rf}} z^r_{kl} \qquad ij \in A^f, r \in R \tag{3}$$

$$d^r_{kl} z^r_{kl} \leq t_i \qquad kl \in A^r, li \in A^{rf}, i \in N^f, r \in R \tag{4}$$

$$\sum_{l \in N^r} z^r_{kl} \leq \bar{x}^r_k \qquad \forall k \in N^r, r \in R \tag{5}$$

$$y_i \geq 1 - t_i/T \qquad i \in N^f \tag{6}$$

$$t_i \geq 0 \qquad i \in N^f \tag{7}$$

$$z^r_{kl} \in \{0,1\} \qquad kl \in A^r, r \in R \tag{8}$$

$$y_i \in \{0,1\} \qquad i \in N^f \tag{9}$$

The objective function (1) minimises the burned area. Constraints (2) define the nodes of the fire initial perimeter, that is, the burning nodes when the fire is detected. Constraints (3) define the fire spread: the fire arrival time at each node is the minimum of the potential fire arrival times from each of the incoming arcs. For a detailed discussion of this type of fire spread modelling, we refer the reader to [10]. A resource in range of the origin of an arc delays the fire transmission from that origin node beyond the time horizon T. Constraints (4) state that a resource can only be moved to a node if the resource manages to reach the node before the fire. Constraints (5) state that a resource can only be moved from a position if it was placed there. There can only be one resource of type r per node. Constraints (6) relate the fire arrival time with the node being burned or not at instant T. The remaining constraints, (7), (8) and (9), define the domain of the decision variables.

Figure 1 illustrates a solution to the deterministic problem.

3 The Two-Stage Distributionally Robust Optimisation Model

Here we introduce a two-stage distributionally robust optimization (DRO) model. The prepositioning of available resources is the first-stage decision, taken before the uncertainty is revealed (i.e., the location of ignitions and wind direction). The initial locations of the resources in the deterministic model of the previous Section are no longer fixed but can be chosen. When the wildfire ignition occurs, resources are then relocated to prevent fire propagation; these are the second-stage decisions. The resulting program is known as a recourse problem, in which certain decisions can be made after the uncertainty is disclosed. To model uncertainty, we consider a set of scenarios $\Omega = \{\omega_1, \ldots, \omega_n\}$.

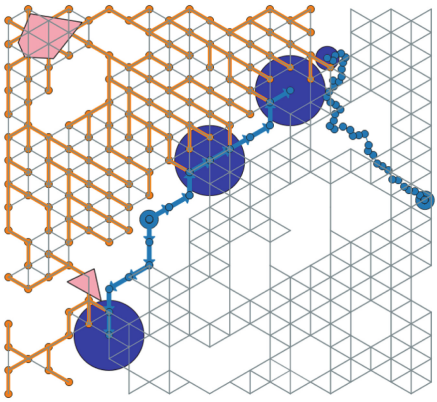


Fig. 1. Illustration of a solution of the deterministic model. Pink polygons are ignition areas. The fire graph is represented by the grey arcs. Orange arcs are arcs in which the fire is effectively transmitted. Blue arcs and nodes correspond to the movement of the resources from their initial positions to their attack positions, in which their attack range is represented by the larger blue circles. All arcs are undirected for clarity. The full resource network is not represented also for clarity. (Color figure online)

The two-stage DRO model can be written as a min-max problem:

$$\min_{x \in X} \left\{ \max_{P \in \mathbb{A}} \sum_{\omega \in \Omega} P_\omega \, Q(x, \omega) \right\},$$

where X is the set of feasible prepositioning locations, $\mathbb{A}$ is the ambiguity set, P_ω is the probability of scenario ω, and $Q(x, \omega)$ represents the recourse cost (i.e., the burned area) given prepositioning decision x and scenario ω.

Next, we discuss in detail the ambiguity set and present the full formulation as a single-stage model.

3.1 The Ambiguity Set

The ambiguity set $\mathbb{A}$ of probability distributions is defined on the finite support Ω. A scenario $\omega \in \Omega$ is represented by the pair $(\alpha_\omega, I_\omega)$, where α_ω is the wind direction and I_ω is the set of ignition locations.

We consider a variant of the Kantorovich ambiguity set which is defined as the set of probability mass distributions that can be obtained from the nominal distribution through a constrained transportation problem.

For a scenario ω represented by $(\alpha_\omega, I_\omega)$ and another scenario ω' represented by $(\alpha_{\omega'}, I_{\omega'})$, let $D^I(\omega, \omega')$ denote the distance between ignitions. This distance is computed using the fire spread model MTT under the average wind conditions between the scenarios. The distance $D(\omega, \omega')$ is defined as:

$$D(\omega, \omega') = a^1 \min\{|\alpha_\omega - \alpha_{\omega'}|, 360 - |\alpha_\omega - \alpha_{\omega'}\}| + a^2 D^I(\omega, \omega'). \quad (10)$$

The coefficients a serve as normalisation values, but may also be assigned user-specific weights. For instance, if we wish to consider only the distance between wind directions, we can set $a^1 = 1$ and $a^2 = 0$.

In this work, we rely only on the maximum value of each distance for normalisation:

$$a^1 = \frac{1}{\max_{\omega,\omega' \in \Omega} \min\{|\alpha_\omega - \alpha_{\omega'}|, 360 - |\alpha_\omega - \alpha_{\omega'}|\}},$$
$$a^2 = \frac{1}{\max_{\omega,\omega' \in \Omega} D^I(\omega, \omega')}.$$

The distance between two probability distributions P and Q is given by:

$$W(P,Q) = \min_{k \in J(P,Q)} \sum_{\omega,\omega' \in \Omega} D(\omega,\omega')\, k_{\omega\omega'},$$

where Q denotes the nominal probability distribution, and $J(P,Q)$ is the set of all couplings (joint probability distributions) with marginals P and Q.

We define our ambiguity set $\mathbb{A}_\epsilon$ such that $W(P,Q) \leq \epsilon$. The ambiguity set is the feasible set of a transportation problem with variables P_ω (probability mass) and $k_{\omega\omega'}$ (transported mass between scenarios). This transportation problem is defined as:

$$\mathbb{A}_\epsilon = \Big\{ \sum_{\omega \in \Omega} \sum_{\omega' \in \Omega} D(\omega,\omega') k_{\omega\omega'} \leq \epsilon, \tag{11}$$
$$\sum_{\omega' \in \Omega} k_{\omega\omega'} = P_\omega, \qquad \omega \in \Omega, \tag{12}$$
$$\sum_{\omega \in \Omega} k_{\omega\omega'} = Q_{\omega'}, \qquad \omega' \in \Omega, \tag{13}$$
$$\sum_{\omega \in \Omega} P_\omega = 1, \tag{14}$$
$$P_\omega \geq 0, \qquad \omega \in \Omega, \tag{15}$$
$$k_{\omega\omega'} \geq 0, \qquad \omega, \omega' \in \Omega \Big\}. \tag{16}$$

Parameter ϵ controls the maximum total transportation amount and, consequently, the size of the ambiguity set. When $\epsilon = 0$, the ambiguity set reduces to the singleton $\{Q\}$ (no transportation), yielding the stochastic program, where a set of scenarios and the probability of their occurrence is known. For sufficiently large ϵ, all probability mass can be transported to a single scenario, resulting in the robust problem (worst-case approach).

The distribution separation problem, a variant of the well-known transportation problem, is:

$$\max_{P \in \mathbb{A}} \sum_{\omega \in \Omega} P_\omega\, Q(x,\omega) \tag{17}$$
$$\text{s.t. } (11) - (16),$$

where x is fixed from the first-stage problem.

3.2 DRO Model

The mixed-integer programming (MIP) model that addresses the robust prepositioning problem includes static (first-stage) decision variables defining the prepositions. The parameters $\bar{x}_k^r$ of the deterministic model of the previous Section become decision variables:

$$x_k^r = \begin{cases} 1, & \text{if a resource of type } r \text{ is prepositioned at node } k, \\ 0, & \text{otherwise,} \end{cases} \quad k \in N^r,\ r \in R.$$

To model uncertainty, we consider a set of scenarios Ω. Each scenario $\omega \in \Omega$ represents a set of ignition locations, wind speed, and direction. For each scenario, we define the following adjustable (second-stage) decision variables:

t_i^ω : fire arrival time at node i in scenario ω, $\quad i \in N^f$;

$$z_{kl}^{\omega r} = \begin{cases} 1, \text{ if a resource of type } r \text{ prepositioned at node } k \\ \quad \text{is moved to node } l \text{ in scenario } \omega, \\ 0, \text{otherwise,} \end{cases} \quad kl \in A^r,\ r \in R;$$

$$y_i^\omega = \begin{cases} 1, & \text{if node } i \text{ burns before time } T \text{ in scenario } \omega, \\ 0, & \text{otherwise,} \end{cases} \quad i \in N^f.$$

To formulate a complete MILP, we use the well-known epigraph reformulation to remove the inner maximization problem. We introduce an auxiliary variable θ representing the worst-case expected burned area:

$$\theta = \max_{P \in \mathbb{A}} \sum_{\omega \in \Omega} P_\omega \left(\sum_{i \in N^f} y_i^\omega \right). \tag{18}$$

The full model is:

$$\min \theta \tag{19}$$

s.t.

$$\theta \geq \sum_{\omega \in \Omega} P_\omega \left(\sum_{i \in N^f} y_i^\omega \right), \qquad P \in \mathbb{A}, \tag{20}$$

$$\sum_{k \in N^r} x_k^r \leq b^r, \qquad r \in R, \tag{21}$$

$$t_i^\omega = 0, \qquad i \in I^f, \omega \in \Omega, \tag{22}$$

$$\sum_{k \in N^r} z_{kl}^{\omega r} \leq 1, \qquad r \in R, \omega \in \Omega, \tag{23}$$

$$t_j^\omega \leq t_i^\omega + c_{ij}^\omega + T \sum_{kl \in A^r : l = i} z_{kl}^{\omega r}, \qquad ij \in A^f, r \in R, \omega \in \Omega, \tag{24}$$

$$d_{kl}^r z_{kl}^{\omega r} \leq t_l^\omega, \qquad kl \in A^r, li \in A^{rf}, r \in R, \omega \in \Omega, \tag{25}$$

$$\sum_{l \in N^r} z_{kl}^{\omega r} = x_k^r, \qquad k \in N^r, r \in R, \omega \in \Omega, \tag{26}$$

$$y_i^\omega \geq 1 - t_i^\omega / T, \qquad i \in N^f, \omega \in \Omega, \tag{27}$$

$$t_i^\omega \geq 0, \qquad i \in N^f, \omega \in \Omega, \tag{28}$$

$$x_k^r \in \{0, 1\}, \qquad k \in N^r, r \in R, \tag{29}$$

$$z_{kl}^{\omega r} \in \{0, 1\}, \qquad kl \in A^r, r \in R, \omega \in \Omega, \tag{30}$$

$$y_i^\omega \in \{0, 1\}, \qquad i \in N^f, \omega \in \Omega. \tag{31}$$

Constraints 20 assure that all the distribution of probabilities of the ambiguity set are taken into account. These constraints, together with the objective function 19, assure the burned area in the worst case distribution of probabilities is minimized.

Note that (20) represents an infinite family of constraints. To solve the model, we use an exact decomposition algorithm that considers only a subset of $\mathbb{A}$. Initially, this subset, denoted by A is set to the singleton $A = \{Q\}$. Then, in each iteration, for a given first-stage solution, we solve the transportation problem to check for violated inequalities; when found, the corresponding probability distribution is added to A.

Constraints (21) limit the number of resource units of each type. Constraints (22) define the nodes in the initial fire perimeter (burning nodes at detection time). Constraints (26) ensure that a resource can only move from a node (second-stage) if it is initially there (first stage).

The following constraints follow the deterministic model directly, but are repeated for each scenario: Constraints (23) limit the number of resources of type r in a node to one, in each scenario. Constraints (24) model fire spread in each scenario. Constraints (25) ensure that a resource can only move to a node if it arrives before the fire in each scenario. Constraints (27) link fire arrival time

with whether a node burns before time T. The remaining constraints define variable domains.

4 Decomposition Algorithm

In [25], several solution approaches were compared to solve DRO problems with finite supports. The study concludes that a solution approach should be selected according to the ambiguity set characteristics. These solution approaches comprise the dualization approach, which is the simplest to apply but may not be suitable to instances that are simultaneously hard and have large number of scenarios; the row-and-column generation generalizes the same approach for Robust Optimization, and it is suitable exactly when the ambiguity set approximates the RO case, and the L-shape is the best approach for the harder instances with many scenarios. As in fire suppression, preventing the worst-case scenarios is critical, we use the row-and-column approach. This algorithm is an exact decomposition method to solve the two-stage DRO model. It iteratively alternates between solving a master problem (where only few scenarios are considered) and a separation problem to identify the worst-case probability distribution. The corresponding cut (20) and the scenarios with positive probability in that cut are added to the master problem. The algorithm is provided in Algorithm 1.

Algorithm 1 Dynamic decomposition algorithm for the DRO model

Initialize $i \leftarrow 0$, $UB \leftarrow \infty$, $LB \leftarrow -\infty$, $Gap \leftarrow \infty$, and $\mathcal{A} \leftarrow \{Q\}$;
Build the initial master problem MR_0 (including only first-stage variables and constraints);
Solve MR_0; obtain x' and LB;
while $Gap > tol$ **and** $i \leq maxiter$ **do**
  Solve the transportation problem (17) using x' and $R(x', \omega)$;
  Let P_ω be the resulting probabilities;
  if $UB > \sum_{\omega \in \Omega} P_\omega R(z', \omega)$ **then**
    $UB \leftarrow \sum_{\omega \in \Omega} P_\omega R(z', \omega)$;
  end if
  $i \leftarrow i + 1$;
  Extend MR_{i-1} to MR_i by adding:

- the inequality $\theta \geq \sum_{\omega \in \Omega} P_\omega R(x', \omega)$,
- for each ω with $P_\omega > 0$, the corresponding second-stage variables $y_i^\omega, z_{kl}^{\omega r}, t_i^\omega$ and constraints (22)–(31).

  Solve MR_i; update x' and LB;
  Compute $Gap = 100 \times (UB - LB)/UB$;
end while

5 Computational Experiments

The computational tests were performed on an 8th Gen Intel i7 CPU running at 1.8 GHz. The optimization models were implemented in Python using the Gurobi solver (version 10.0.2) and pyO3F-a Python framework developed within the project "An optimization framework for reducing forest fire" [27,28]. This framework enables the definition of fire and resource networks from an actual landscape, given a specified resolution and fire spread scenarios that depend on fuels, slopes, and wind conditions.

The experiments were conducted on a landscape surrounding Baião, a municipality in northern Portugal. The study area spans approximately 20 km by 18 km (around 360 km^2). A resolution of 1000 m was used, corresponding to the distance between adjacent nodes in the hexagonal grid representing the landscape.

For all tests, the initial fire perimeters were obtained by simulating fire spread during 30 min, and the time horizon was set to 120 min.

Only one type of resource was considered, with speed varying according to terrain: from 5 km/h on paths unsuitable for vehicles to from 60 km/h on primary and secondary roads (using the Open Street map categorizations). Each resource unit had a suppression radius of 1000 m, meaning it suppresses fire on any fire node within this distance.

For each instance, a preprocessing phase was conducted reducing the number of nodes in the fire and resource networks. In the fire network we remove every node whose quickest path from the ignition is greater than T. For the resource network, we created a grid and kept only the nodes closest to the intersections.

5.1 Instances and Scenarios Generation

The classification of instances is based on the assumption that relevant information, such as weather forecasts and estimated ignition counts, is available on the day prior to operations. This enables a preliminary assessment of the risk level for the upcoming day. Instances are defined according to these two key parameters, wind intensity and the expected number of ignitions, and on the number of available resources.

Six instance types are identified as follows:

1. **High Risk - Very Dangerous**
   - Wind speed: 60 km/h (very strong)
   - Number of ignitions: 4 (high)
   - Instance set: `60_4`
2. **Moderate-to-High Risk - Dangerous**
   - Wind speed: 60 km/h (very strong)
   - Number of ignitions: 1 or 2 (low)
   - Instance sets: `60_1` and `60_2`
3. **Moderate-to-High Risk - Dangerous**
   - Wind speed: 30 km/h (strong)
   - Number of ignitions: 4 (high)

   - Instance set: `30_4`
4. **Moderate Risk - Less Dangerous**
   - Wind speed: 30 km/h (strong)
   - Number of ignitions: 1 or 2 (low)
   - Instance sets: `30_1` and `30_2`

For each set of instances we generate 80 scenarios, that differ on the wind direction and on the ignition location(s):

- For instances with one ignition (labelled `60_1` and `30_1`): four wind directions and 20 combinations of two ignition locations.
- For instances with two ignitions (labelled `60_2` and `30_2`): four wind directions and 20 combinations of two ignition locations.
- For instances with four ignitions (labelled `60_4` and `30_4`): four wind directions and 20 combinations of four ignition locations.

5.2 Computational Results

Table 1 and Table 2 summarise the computational results obtained with Algorithm 1 for all instance sets under wind speeds of 30 km/h and 60 km/h, respectively, and different values of ϵ. Each instance corresponds to a combination of wind direction, number of ignitions and number of available resources, with increasing levels of risk and uncertainty captured by ϵ. The tables present the same information. Column "Inst" gives the name of the instance. The instance is identified by the wind speed, number of ignitions and number of resources. Hence, the case where no resources are employed amounts to a fire spread simulation. Column "ϵ" specifies the parameter used to control the size of the ambiguity set. Columns "LB" and "UB" give the lower and upper bound obtained by Algorithm 1, respectively, while column "Gap" computes the corresponding gap $Gap = 100 \times (UB - LB)/UB$. Column "$\#S$" gives the number of scenarios added. The three columns below Time give the running time, in sec., spent with the master problem (column "M"), subproblem (column "S") and the transportation problem (column "T"). Finally, the number of iterations is given in column "Iter".

Overall, the model achieved zero optimality gaps across nearly all experiments, indicating that the solution approach consistently identified optimal or near-optimal solutions within acceptable computational limits. This is particularly relevant given the complexity of the problem.

For the instances with wind speed 30 km/h (Table 1), the runtimes (sum of columns M, S, and T) ranged from a few seconds up to a total of more than 4000 sec. in the most demanding case with 8 iterations. The hardest instances are those with one ignition and one resource. Increasing the number of resources makes such instances easier. As ϵ increases, fewer scenarios are activated (see $\#S$ column) since, as explained above, the probability mass concentrates in fewer scenarios. In all cases the number of iteration is very small.

Table 1. Results of the dynamic DRO model of the instances with wind speed 30 km/h.

						Runtime			
Inst	ϵ	LB	UB	Gap	$\#S$	M	S	T	Iter
30_1_1	0	5.585	5.585	0.0	31	682	101	17	4
30_1_1	0.1	6.678	6.678	0.0	18	541	111	19	4
30_1_1	0.2	7.0	7.0	0.0	15	3811	244	37	8
30_1_1	0.3	7.0	7.0	0.0	13	1859	190	32	7
30_1_1	0.4	7.0	7.0	0.0	12	1740	186	31	7
30_1_1	0.5	7.0	7.0	0.0	12	1157	139	24	6
30_1_3	0	5.368	5.368	0.0	21	13	27	5	1
30_1_3	0.1	6.638	6.638	0.0	7	5	37	5	1
30_1_3	0.2	7.0	7.0	0.0	3	1	26	4	1
30_1_3	0.3	7.0	7.0	0.0	3	2	30	4	1
30_1_3	0.4	7.0	7.0	0.0	3	2	33	5	1
30_1_3	0.5	7.0	7.0	0.0	3	1	24	4	1
30_1_10	0	5.368	5.368	0.0	21	11	26	4	1
30_1_10	0.1	6.638	6.638	0.0	7	4	34	5	1
30_1_10	0.2	7.0	7.0	0.0	3	2	29	5	1
30_1_10	0.3	7.0	7.0	0.0	3	2	28	5	1
30_1_10	0.4	7.0	7.0	0.0	3	1	26	4	1
30_1_10	0.5	7.0	7.0	0.0	3	2	32	5	1
30_2_1	0	12.187	12.187	0.03	24	713	118	16	3
30_2_1	0.1	15.2	15.2	0.0	5	124	121	17	3
30_2_1	0.2	16.0	16.0	0.0	2	15	75	12	2
30_2_1	0.3	16.0	16.0	0.0	2	16	89	12	2
30_2_1	0.4	16.0	16.0	0.0	2	12	60	10	2
30_2_1	0.5	16.0	16.0	0.0	2	14	67	11	2
30_2_3	0	10.553	10.553	0.0	22	15	35	5	1
30_2_3	0.1	12.408	12.408	0.0	10	8	43	6	1
30_2_3	0.2	13.0	13.0	0.0	3	3	41	6	1
30_2_3	0.3	13.0	13.0	0.0	3	3	44	6	1
30_2_3	0.4	13.0	13.0	0.0	3	2	34	5	1
30_2_3	0.5	13.0	13.0	0.0	3	2	36	5	1
30_2_10	0	10.553	10.553	0.0	22	19	38	5	1
30_2_10	0.1	12.408	12.408	0.0	10	10	50	7	1
30_2_10	0.2	13.0	13.0	0.0	3	3	53	6	1
30_2_10	0.3	13.0	13.0	0.0	3	2	42	6	1
30_2_10	0.4	13.0	13.0	0.0	3	2	36	5	1
30_2_10	0.5	13.0	13.0	0.0	3	2	30	5	1
30_4_1	0	25.396	25.396	0.0	22	486	84	26	2
30_4_1	0.1	29.915	29.915	0.0	11	95	81	25	2
30_4_1	0.2	31.637	31.637	0.0	6	56	84	25	2
30_4_1	0.3	32.0	32.0	0.0	2	20	95	28	2
30_4_1	0.4	32.0	32.0	0.0	2	17	95	26	2
30_4_1	0.5	32.0	32.0	0.0	2	23	110	29	2
30_4_3	0	20.222	20.23	0.039	24	71	87	25	2
30_4_3	0.1	23.484	23.484	0.0	11	27	93	26	2
30_4_3	0.2	25.0	25.0	0.0	3	9	98	27	2
30_4_3	0.3	25.0	25.0	0.0	3	5	80	24	2
30_4_3	0.4	25.0	25.0	0.0	3	6	110	29	2
30_4_3	0.5	25.0	25.0	0.0	3	8	121	33	2
30_4_10	0	19.976	19.976	0.0	22	20	47	13	1
30_4_10	0.1	23.285	23.285	0.0	10	9	43	12	1
30_4_10	0.2	24.927	24.927	0.0	3	4	44	13	1
30_4_10	0.3	25.0	25.0	0.0	2	2	48	13	1
30_4_10	0.4	25.0	25.0	0.0	2	3	50	14	1
30_4_10	0.5	25.0	25.0	0.0	2	3	52	14	1

Table 2. Results of the dynamic DRO model of the instances with wind speed 60 km/h.

						Runtime			
Inst	ϵ	LB	UB	Gap	$\#S$	M	S	T	Iter
60_1_1	0	4.379	4.233	0.0	24	19	23	4	1
60_1_1	0.1	5.0	5.0	0.0	8	4	23	4	1
60_1_1	0.2	5.0	5.0	0.0	8	4	24	4	1
60_1_1	0.3	5.0	5.0	0.0	8	4	22	4	1
60_1_1	0.4	5.0	5.0	0.0	8	10	52	5	1
60_1_1	0.5	5.0	5.0	0.0	8	4	23	4	1
60_1_3	0	4.241	4.233	0.202	24	12	24	4	1
60_1_3	0.1	5.0	5.0	0.0	8	3	22	4	1
60_1_3	0.2	5.0	5.0	0.0	8	3	23	4	1
60_1_3	0.3	5.0	5.0	0.0	8	3	39	4	1
60_1_3	0.4	5.0	5.0	0.0	8	7	34	7	1
60_1_3	0.5	5.0	5.0	0.0	8	3	22	4	1
60_1_10	0	4.233	4.233	0.0	24	9	23	4	1
60_1_10	0.1	5.0	5.0	0.0	8	3	23	4	1
60_1_10	0.2	5.0	5.0	0.0	8	3	22	4	1
60_1_10	0.3	5.0	5.0	0.0	8	5	29	5	1
60_1_10	0.4	5.0	5.0	0.0	8	3	24	4	1
60_1_10	0.5	5.0	5.0	0.0	8	3	22	4	1
60_2_1	0	9.431	9.431	0.0	26	240	58	9	2
60_2_1	0.1	11.0	11.0	0.0	6	41	74	11	2
60_2_1	0.2	11.0	11.0	0.0	5	38	96	15	3
60_2_1	0.3	11.0	11.0	0.0	5	44	97	14	3
60_2_1	0.4	11.0	11.0	0.0	5	32	78	13	3
60_2_1	0.5	11.0	11.0	0.0	5	31	73	13	3
60_2_3	0	8.387	8.387	0.0	22	13	28	4	1
60_2_3	0.1	9.744	9.744	0.0	8	5	36	5	1
60_2_3	0.2	10.0	10.0	0.0	3	1	34	5	1
60_2_3	0.3	10.0	10.0	0.0	3	3	47	7	1
60_2_3	0.4	10.0	10.0	0.0	3	1	27	4	1
60_2_3	0.5	10.0	10.0	0.0	3	1	31	4	1
60_2_10	0	8.387	8.387	0.0	22	15	35	5	1
60_2_10	0.1	9.744	9.744	0.0	8	6	41	5	1
60_2_10	0.2	10.0	10.0	0.0	3	1	33	5	1
60_2_10	0.3	10.0	10.0	0.0	3	1	30	4	1
60_2_10	0.4	10.0	10.0	0.0	3	1	27	4	1
60_2_10	0.5	10.0	10.0	0.0	3	1	27	4	1
60_4_1	0	18.935	18.935	0.0	23	83	32	12	1
60_4_1	0.1	21.714	21.714	0.0	8	19	37	12	1
60_4_1	0.2	22.0	22.0	0.0	5	8	31	12	1
60_4_1	0.3	22.0	22.0	0.0	5	11	35	12	1
60_4_1	0.4	22.0	22.0	0.0	5	7	30	12	1
60_4_1	0.5	22.0	22.0	0.0	5	8	32	12	1
60_4_3	0	16.086	16.094	0.054	23	74	67	24	2
60_4_3	0.1	18.767	18.767	0	9	36	133	47	4
60_4_3	0.2	19.0	19.0	0.0	5	8	69	24	2
60_4_3	0.3	19.0	19.0	0.0	5	12	82	26	2
60_4_3	0.4	19.0	19.0	0.0	5	10	74	24	2
60_4_3	0.5	19.0	19.0	0.0	5	11	84	26	2
60_4_10	0	15.504	15.504	0.0	22	14	34	12	1
60_4_10	0.1	18.498	18.498	0.0	7	5	34	12	1
60_4_10	0.2	19.0	19.0	0.0	2	4	38	22	1
60_4_10	0.3	19.0	19.0	0.0	2	1	42	13	1
60_4_10	0.4	19.0	19.0	0.0	2	1	34	12	1
60_4_10	0.5	19.0	19.0	0.0	2	2	36	12	1

For the instances with wind speed 60 km/h (Table 2), the computational times are generally smaller and more stable than in the 30 km/h case. Most instances were solved within a few seconds, with only a few reaching higher runtimes, particularly those involving four ignitions and three resources, where up to four iterations were required. As with the previous case, increasing the number of resources tends to make the instances easier to solve, and the number of activated scenarios (#S) decreases as ϵ grows, confirming the expected concentration of probability mass in fewer representative scenarios. Optimality gaps remain equal to zero in all cases, showing that the solution method maintains its robustness and efficiency under higher wind speed conditions.

Overall, the dynamic DRO model demonstrated strong computational performance and stability across all tested configurations, requiring only a few iterations to converge. The proposed approach efficiently handles different levels of uncertainty, while reaching optimality within reasonable computational effort.

From the operational point of view, we can observe that in most instances increasing the number of resources has no meaningful impact. For example, in instances 30_1 employing more than one resource has no impact for the robust case, i.e., for $\epsilon = 0$. While in the worst-case of four ignition, case 30_4, there is a reduction on the burned area when 3 resources are employed instead of 1, but there is no further improvement by increasing the number of resources.

6 Conclusions

This work proposed a two-stage Distributionally Robust Optimisation model for the prepositioning of firefighting resources under uncertain ignition locations and wind conditions. The approach incorporates an ambiguity set based on the Wasserstein metric, allowing the model to hedge against uncertain weather conditions and location of ignitions. To efficiently solve the resulting problem, we developed an exact decomposition algorithm that generates second-stage variables and constraints only for scenarios with positive probability, significantly improving the scalability of the instances solved.

The computational experiments demonstrated that the proposed algorithm is computationally efficient by solving the tested instances to optimality or near optimality, always using a small number of iterations, confirm the suitability of the proposed approach for addressing complex wildfire resource allocation problems under uncertainty.

From an operational perspective, the DRO approach provides stable and interpretable solutions that remain effective under a wide range of uncertainty realisations. Moreover, it provides insights not only on the place to allocate the resources but also on the impact on the number of resources on the fire suppression.

Acknowledgements. The work of the first author was supported by CEMS.UL - Center for Mathematical Studies (https://doi.org/10.54499/UID/04561/2025) under the Portuguese Foundation for Science and Technology Multi-Annual Financing Program for R&D Units, grant UID/04561/2025.

The work of the second author was supported by FCT – Portuguese Foundation for Science and Technology within the R&D Unit

Project Scope UID/00319/2025 - Centro ALGORITMI (ALGORITMI/UM, https://doi.org/10.54499/UID/00319/2025) and by the Firesys (AI-empowered Decision Support System for Wildfire Management) project, funded by the National Innovation Agency, Portugal 2030, Innovation and Digital Transition Programme, Compete 2030 (project 21378).

This work was also supported by CIDMA (https://ror.org/05pm2mw36) under the Portuguese Foundation for Science and technology (FCT, https://ror.org/00snfqn58), Grants UID/04106/2025 (https://doi.org/10.54499/UID/04106/2025) and UID/PRR/04106/2025 (https://doi.org/10.54499/UID/PRR/04106/2025).

References

1. Martell, D.L.: A review of recent forest and wildland fire management decision support systems research. Curr. Forestry Rep. **1**, 128–137 (2015)
2. Martell, D.L.: A review of operational research studies in forest fire management. Can. J. For. Res. **12**(2), 119–140 (1982)
3. Mendes, A.B., Alvelos, F.P.: Iterated local search for the placement of wildland fire suppression resources. Eur. J. Oper. Res. **304**(3), 887–900 (2023)
4. Granda, B., León, J., Vitoriano, B., Hearne, J.: Decision support models and methodologies for fire suppression. Fire **6**(2) (2023)
5. Hof, J., Omi, P.N., Bevers, M., Laven, R.D.: A timing-oriented approach to spatial allocation of fire management effort. Forest Sci. **46**(3), 442–451 (2000)
6. Finney, M.A.: Fire growth using minimum travel time methods. Can. J. For. Res. **32**(8), 1420–1424 (2002)
7. Finney, M.A., et al.: An overview of FlamMap fire modeling capabilities. In: Fuels Management—How to Measure Success: Conference Proceedings, USDA Forest Service (2006)
8. Wei, Y., Rideout, D.B., Hall, T.B.: Toward efficient management of large fires: a mixed integer programming model and two iterative approaches. Forest Sci. **57**(5), 435–447 (2011)
9. Belval, E.J., Wei, Y., Bevers, M.: A mixed integer program to model spatial wildfire behavior and suppression placement decisions. Can. J. For. Res. **45**(4), 384–393 (2015)
10. Alvelos, F.: Mixed integer programming models for fire fighting. In: Gervasi, O., et al. (eds.) ICCSA 2018. LNCS, vol. 10961, pp. 637–652. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95165-2_45
11. Harris, M.G., Forbes, M.A., Taimre, T.: Logic-based Benders decomposition for wildfire suppression. Comput. Oper. Res., 106392 (2023)
12. Alvelos, F., Marto, M., Mendes, A.: Dispatching, positioning and routing resources for wildfire initial attack. Networks **86**(1), 80–102 (2025)
13. Hu, X., Ntaimo, L.: Integrated simulation and optimization for wildfire containment. ACM Trans. Model. Comput. Simul. **19**(4), 1–29 (2009)
14. Ntaimo, L., Gallego Arrubla, J.A., Stripling, C., Young, J., Spencer, T.: A stochastic programming standard response model for wildfire initial attack planning. Can. J. Forest Res. **42**(6), 987–1001 (2012)

15. Lee, Y., Fried, J.S., Albers, H.J., Haight, R.G.: Deploying initial attack resources for wildfire suppression: spatial coordination, budget constraints, and capacity constraints. Can. J. For. Res. **43**(1), 56–65 (2013)
16. Belval, E.J., Wei, Y., Bevers, M.: A stochastic mixed integer program to model spatial wildfire behavior and suppression placement decisions with uncertain weather. Can. J. For. Res. **46**(2), 234–248 (2016)
17. Belval, E.J., Wei, Y., Bevers, M.: Modeling ground firefighting resource activities to manage risk given uncertain weather. Forests **10**(12), 1077 (2019)
18. Mendes, A.B., Alvelos, F.P.: A robust optimisation approach for the placement of forest fire suppression resources. Int. Trans. Oper. Res. **32**(3), 1312–1342 (2025)
19. Marques, F., Agra, A., Alvelos, H., Xambre, A.R., Alvelos, F.: A robust approach for the prepositioning of resources for wildfire suppression. In: Alvelos, F., Martins, I., Rocha, A.M.A.C. (eds) Advances in Optimization and Wildfire. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-03108-2_10
20. Bansal, M., Huang, K.-L., Mehrotra, S.: Decomposition algorithms for two-stage distributionally robust mixed binary programs. SIAM J. Optim. **28**(3), 2360–2383 (2018)
21. Lin, F., Fang, X., Gao, Z.: Distributionally robust optimization: a review on theory and applications. Numer. Algebra, Control Optim. **12**(1), 159–212 (2022)
22. Rahimian, H., Mehrotra, S.: Frameworks and results in distributionally robust optimization. Open J. Math. Optim. **3** (2022)
23. Agra, A., Cerveira, A.: Wind farm layout optimization under uncertainty. TOP **32**, 202–223 (2024)
24. Alvelos, H., Xambre, A.R., Marques, F., Agra, A., Alvelos, F.: Modelling wind behaviour for the development of scenarios in the context of wildfire spread. In F. Alvelos, I. Martins, and A. Rocha (eds.), Advances in Optimization and Wildfire: Proceedings of the First Optimization and Wildfire Conference, Springer, Cham (2026). https://doi.org/10.1007/978-3-032-03108-2_7
25. Agra, A.: Two-stage distributionally robust optimization with a finite support. Comput. Oper. Res. **183**, 107142 (2025)
26. Rothermel, R.C.: A mathematical model for predicting fire spread in wildland fuels. Intermountain Forest and Range Experiment Station, Forest Service, USA (1972)
27. Alvelos, F.: pyO3F – a python framework for wildfire-related optimization, Part I: Design and fundamentals. In: Alvelos, F., Martins, I., Rocha, A. (eds.), Advances in Optimization and Wildfire: Proceedings of the First Optimization and Wildfire Conference. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-03108-2_3
28. Alvelos, F.: pyO3F – a python framework for wildfire-related optimization, Part II: usage. In: Alvelos, F., Martins, I., Rocha, A. (eds.), Advances in Optimization and Wildfire: Proceedings of the First Optimization and Wildfire Conference, Springer, Cham (2026). https://doi.org/10.1007/978-3-032-03108-2_4

A Convergence Analysis of a 1-Norm Neighborhood Mehrotra-Type Algorithm

A. Teixeira[1,2(✉)] and R. De Almeida[1,2]

[1] Department of Mathematics University of Trás-os-Montes e Alto Douro, 5000-801 Vila Real, Portugal

[2] Centro de Matemática, Universidade do Minho - Polo CMAT-UTAD, Braga, Portugal

{ateixeir,ralmeida}@utad.pt

Abstract. Interior-point methods are a class of methods that explore the interior of the feasible region of a problem aiming to reach the optimal solution, contrasting with the simplex method that pursues the same goal by moving, in each iteration, from an extreme point to an adjacent extreme point. In this work, the complexity of a Linear Programming feasible predictor-corrector interior-point variant of the Mehrotra's algorithm using the 1-norm neighborhood is studied. It is proven that the algorithm has polynomial complexity and Q-linear convergence. To achieve these conclusions, some secondary lemmas and propositions are developed and demonstrated. In these results upper bound estimates of both the non linear term in the predictor direction and the non linear term in corrector direction are obtained. Additionally, during the procedure, step size bounds are derived and both the duality gap reduction and the complexity bound are obtained. Within this process, it is proven that, at each iteration, the step size of this 1-norm neighborhood algorithm is bounded by a function depending on the neighborhood parameter and on the dimension of the problem, thus, establishing that it has iteration complexity of $O\left(n^{\frac{3}{2}}\left|\log(\epsilon)\right|\right)$.

Keywords: Feasible Mehrotra-type algorithm · Convergence · 1-norm neighborhood

1 Introduction

Interior-point methods (IPM), due to Karmarkar [9], have become essential for solving large-scale Linear Programming (LP) problems, offering a robust and efficient alternative to traditional simplex methods [5]. It is worth to recall that while IPM explore the interior of the feasible region of a problem aiming to reach the optimal solution, the simplex method pursues the same goal by moving, in each iteration, from an extreme point to an adjacent extreme point.

One of the IPM variants are the predictor-corrector (PC) methods, that combine a predictor step, which approximates the central path, with a corrector

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 301–313, 2027.
https://doi.org/10.1007/978-3-032-30521-3_19

step, which restores centrality. The performance of these methods depends on the definition of the central path's neighborhood, which determines how far the algorithm can deviate from centrality without compromising convergence.

The 2-norm neighborhood defines an ellipsoidal region, strongly penalizing individual deviations in the complementary product; thus, forcing the algorithm to maintain a rigid proximity to the central path, even when such a restriction is not necessary to ensure feasibility or progress. The infinity norm neighborhood represents the opposite extreme. By controlling only the largest individual deviation, it creates a highly permissive hypercubic region, allowing some components to deviate significantly from the mean. Although flexibility can be advantageous (in very elongated predictive directions) it can also require stronger corrections. The infinity norm offers maximum directional freedom, but with less overall control over the algorithm's behavior. The 1-norm neighborhood can be an intermediate approach, because it defines a polyhedral region that allows to compensate deviations; some components may deviate further from the mean as long as others remain close to it. Thus, it is effective for degenerate problems, instances where certain variables rapidly approach zero, and situations where the rigidity of the 2-norm limits progress. The polyhedral geometry reduces sensitivity to isolated extreme values and allows for larger prediction steps, often resulting in greater numerical robustness. Considering this, using the 1-norm in the central path neighborhood is a strategy that balances directional flexibility, centrality control, and numerical robustness (qualities essential for addressing the challenges of large-scale problems that drive current research in optimization) and attempting to find algorithms that are more stable and better adapted to the characteristics of real-world problems [15,20].

Mehrotra [13] developed one of the most successful PC IPM implemented, both theoretically and commercially. Several variants of this algorithm have been implemented and studied, namely [1,2,12,14,16,17,21,23,25]. Additionally, some works analyze the complexity of LP algorithms, proving their polynomiality, for e.g., [4,6,8,10,11,19,22,24].

In [3] some feasible PC variants of the Mehrotra's LP algorithm applied to transportation and assignment problems are presented; the special structure of the constraint matrix of the problems is taken into consideration. In [2,18] the authors proved the polynomial complexity of the primal-dual PC methodology using the infinity norm and the 2-norm neighborhood, respectively.

With the aim of obtaining an algorithm that balances the smoothness of the 2-norm and the permissiveness of infinity norm, in the present work, a feasible variant of the algorithm in [3], is adapted to use the 1-norm neighborhood and, then, theoretically analyzed. Unlike the neighborhoods based on the 2-norm and the infinite norm, which have been extensively studied in the classical literature on IPM, the 1-norm neighborhood exhibits a polyhedral geometry whose detailed analysis has received less attention. The results obtained here allow for a rigorous clarification of the trade-off between directional flexibility and control of centrality, establishing convergence conditions and complexity bounds that, to the best of our knowledge, have not been systematically developed for

this class of neighborhoods. It is demonstrated that, under assumptions analogous to those considered for Euclidean and infinity norm neighborhoods, the algorithm has polynomial complexity and Q-linear convergence. The step-size bounds derivation, the duality gap reduction, and the complexity bound are also presented, as well as the proofs of some necessary secondary results. This process allowed to prove that the algorithm with the 1-norm neighborhood has iteration complexity of $O\left(n^{\frac{3}{2}}|\log(\epsilon)|\right)$. Additionally, its polyhedral geometry proves particularly well-suited for addressing degenerate situations and large-scale problems, offering a theoretical framework that complements traditional approaches.

In this paper, the algorithm and the fundamentals needed in the remaining of the paper are explained in Sect. 2; the convergence analysis of the algorithm is presented in Sect. 3. Finally, the conclusions are drawn in the last section.

2 Preliminaries

The fundamental concepts and notation used throughout this paper will be presented in this section.

Let A be a real matrix $m \times n$ and b, c real column vectors of order m and n, respectively. The sets of primal and dual feasible solutions are presented, respectively, by

$$\mathbb{F}_P = \{x \in \mathbb{R}^n : \ Ax = b, \ x \geqslant 0\},$$
$$\mathbb{F}_D = \left\{(u, v) \in \mathbb{R}^m \times \mathbb{R}^n : \ A^T u + v = c, \ v \geqslant 0\right\}.$$

The LP primal problem can be written as

$$\text{(P)} \ \min_{x \in \mathbb{F}_P} \ c^T x,$$

and the dual problem as

$$\text{(D)} \ \max_{(u,v) \in \mathbb{F}_D} \ b^T u.$$

Any strictly feasible solution of this pair of problems belongs to the set

$$\mathbb{F}^\circ = \{(x, u, v) \in \mathbb{F}_P \times \mathbb{F}_D : x \neq 0, \ v \neq 0\}.$$

Here, the complexity of a feasible variant of the predictor-corrector algorithm developed by Bastos and Paixão [3], *Algorithm BP*, will be analyzed. The algorithm begins with the selection of the parameters, $\varepsilon > 0$, $0 < \tau \leqslant 1$ and $0 \leqslant \theta < 1$, as well as an initial point (x^0, u^0, v^0) within the 1-norm neighborhood defined in (1), where e is the column vector of ones,

$$\mathbb{V}_1(\theta) = \left\{(x, u, v) \in \mathbb{F}^\circ : \ \|XVe - \tau\mu e\|_1 \leqslant \tau\theta\mu\right\}. \tag{1}$$

Then, to test if an optimal solution was achieved, the duality gap, given by

$$\mu = \frac{c^T x - b^T u}{\beta(n)}, \tag{2}$$

where $\beta(n) = n^{\frac{3}{2}}$ if $n \in]5000, +\infty[$ and $\beta(n) = n^2$ otherwise, is compared with the chosen parameter ε. If the duality gap is inferior to ε, then the algorithm stops and an optimal primal-dual solution is attained. Else, to obtain an optimal primal-dual solution, while the duality gap is not inferior to ε the following computations are performed, where $D_Y = diag(y)$ represents the diagonal matrix of vector y.

1. The predictor direction $(\Delta x^a, \Delta u^a, \Delta v^a)$ that satisfies

$$\begin{cases} AD_{\triangle X}e = 0 \\ A^T D_{\triangle U}e + D_{\triangle V}e = 0 \\ D_V D_{\triangle X}e + D_X D_{\triangle V}e = \tau\mu e - D_X D_V \end{cases} . \tag{3}$$

2. The corrector direction $(\Delta x, \Delta u, \Delta v)$ that satisfies

$$\begin{cases} AD_{\triangle X}e = 0 \\ A^T D_{\triangle U}e + D_{\triangle V}e = 0 \\ D_V D_{\triangle X}e + D_X D_{\triangle V}e = \tau\mu e - D_X D_V - D_{\triangle X^a} D_{\triangle V^a} e \end{cases} . \tag{4}$$

3. The primal-dual step lengths, (α_p, α_d), as the pair of the maximum positive value components, α_1 and α_2, such that

$$(x + \alpha_1 \Delta x,\ u + \alpha_2 \Delta u,\ v + \alpha_2 \Delta v) \in \mathbb{V}_1(\theta).$$

4. The next iterate

$$(x^+, u^+, v^+) = (x, u, v) + 0.9995(\alpha_p \triangle x, \alpha_d \triangle u, \alpha_d \triangle v).$$

In this paper, the following index sets, concerning the components of the predictor direction, are also considered

$$\mathbb{I} = \{1, 2, \cdots, n\},$$
$$\mathbb{I}_+ = \{i \in \mathbb{I} : \triangle x_i^a \triangle v_i^a > 0\}, \qquad \mathbb{I}_- = \{i \in \mathbb{I} : \triangle x_i^a \triangle v_i^a < 0\}.$$

In order to study the convergence of *Algorithm BP*, some results are presented and proved in the next section.

3 Convergence Analysis

An upper bound for the non linear term involving the predictor direction used by *Algorithm BP* is obtained in Lemma 1 and an upper bound for the norm concerning the non linear term of the corrector direction of this algorithm is given by Proposition 1. Moreover, to make this work self contained, some previously published results are also presented, namely Lemmas 2, 3 and 4.

Lemma 1. *Let* $(\triangle x^a, \triangle u^a, \triangle v^a)$ *be the predictor direction of Algorithm BP. Then, for all* $i \in \mathbb{I}_+$, $\theta \in [0, 1[$ *and* $\tau \in]0, 1]$,

$$\triangle x_i^a \triangle v_i^a \leqslant \frac{1}{4} \frac{\theta^2}{(1-\theta)} \tau\mu.$$

Proof. From the third condition of (3), one gets

$$\frac{\triangle x_i^a}{x_i} + \frac{\triangle v_i^a}{v_i} = \frac{\tau\mu}{x_i v_i} - 1. \tag{5}$$

Since

$$0 \leqslant \left(\frac{\triangle x_i^a}{x_i} - \frac{\triangle v_i^a}{v_i}\right)^2 = \left(\frac{\triangle x_i^a}{x_i}\right)^2 + \left(\frac{\triangle v_i^a}{v_i}\right)^2 - 2\frac{\triangle x_i^a \triangle v_i^a}{x_i v_i},$$

then

$$\left(\frac{\triangle x_i^a}{x_i}\right)^2 + \left(\frac{\triangle v_i^a}{v_i}\right)^2 \geqslant 2\frac{\triangle x_i^a \triangle v_i^a}{x_i v_i}. \tag{6}$$

Therefore, using

$$\left(\frac{\triangle x_i^a}{x_i} + \frac{\triangle v_i^a}{v_i}\right)^2 = \left(\frac{\triangle x_i^a}{x_i}\right)^2 + \left(\frac{\triangle v_i^a}{v_i}\right)^2 + 2\frac{\triangle x_i^a \triangle v_i^a}{x_i v_i}$$

$$\geqslant 4\frac{\triangle x_i^a \triangle v_i^a}{x_i v_i}$$

equality (5) and inequality (6), one has

$$\frac{\triangle x_i^a \triangle v_i^a}{x_i v_i} \leqslant \frac{1}{4}\left(\frac{\tau\mu}{x_i v_i} - 1\right)^2$$

$$= \frac{1}{4}\left(\frac{\tau\mu - x_i v_i}{x_i v_i}\right)^2.$$

Since

$$|x_i v_i - \tau\mu| \leqslant \sum_{i=1}^{n} |x_i v_i - \tau\mu| = \|XVe - \tau\mu e\|_1 \leqslant \tau\theta\mu,$$

then follows

$$\triangle x_i^a \triangle v_i^a \leqslant \frac{1}{4}\frac{|\tau\mu - x_i v_i||\tau\mu - x_i v_i|}{x_i v_i} \leqslant \frac{1}{4}\frac{(\tau\theta\mu)^2}{x_i v_i}.$$

Due to

$$(1-\theta)\tau\mu \leqslant x_i v_i \leqslant (1+\theta)\tau\mu$$

one gets

$$\triangle x_i^a \triangle v_i^a \leqslant \frac{1}{4}\frac{(\tau\theta\mu)^2}{x_i v_i} \leqslant \frac{1}{4}\frac{\theta^2}{(1-\theta)}\tau\mu. \tag{7}$$

Henceforth, the condition $\triangle x^{aT}\triangle v^a \geqslant 0$ will be considered for the predictor direction of *Algorithm BP*.

In order to prove Proposition 1 the following result (transcribed from [20, Lemma 5.3 (p.88)] and herein called Lemma 2), presenting an upper bound estimate for $\|D_U D_V e\|_2$, will also be needed.

Lemma 2. *Let $u, v \in \mathbb{R}^n$ with $u^T v \geqslant 0$. Then,*

$$\|D_U D_V e\|_2 \leqslant 2^{-\frac{3}{2}} \|u+v\|_2^2.$$

An upper bound for the norm of the non linear term of the corrector direction of *Algorithm BP* is derived in the next result.

Proposition 1. *Let $(x,u,v) \in \mathbb{V}_1(\theta)$ be the current iterate of Algorithm BP and $(\triangle x, \triangle u, \triangle v)$ be the corrector direction of Algorithm BP. Then,*

$$\|D_{\triangle V} D_{\triangle X} e\|_1 \leqslant C(\theta)\, \tau\mu,$$

for $\triangle x^{aT}\triangle v^a \geqslant 0$, and $C(\theta) = \dfrac{n^{\frac{3}{2}}}{18\sqrt{2}(1-\theta)^3}$.

Proof. Lets start by multiplying the third equation of (4) by $(D_X D_V)^{-\frac{1}{2}}$ and, thus, obtain

$$\begin{aligned}(D_X)^{-\frac{1}{2}}(D_V)^{\frac{1}{2}}\triangle x + (D_V)^{-\frac{1}{2}}(D_X)^{\frac{1}{2}}\triangle v = \\ = (D_X D_V)^{-\frac{1}{2}}\left(\tau\mu^k e - D_{\triangle X^a} D_{\triangle V^a} e - D_X D_V e\right).\end{aligned} \tag{8}$$

Taking the first term of (8) and considering

$$M = (D_V)^{-\frac{1}{2}}(D_X)^{\frac{1}{2}},$$

one has

$$(D_X)^{-\frac{1}{2}}(D_V)^{\frac{1}{2}}\triangle x + (D_V)^{-\frac{1}{2}}(D_X)^{\frac{1}{2}}\triangle \mathrm{v} = M^{-1}\triangle x + M\triangle v.$$

Now, let

$$w_1 = M^{-1}\triangle x, \quad w_2 = M\triangle v, \quad W_1 = \mathrm{diag}(w_1), \quad W_2 = \mathrm{diag}(w_2).$$

Using Lemma 2 and the fact that $\|\cdot\|_2 \leqslant \|\cdot\|_1 \leqslant \sqrt{n}\,\|\cdot\|_2$, follows

$$\begin{aligned}
\|W_1W_2e\|_1 &= \|D_{\triangle V}D_{\triangle X}e\|_1 \\
&\leqslant \sqrt{n}\ \|D_{\triangle V}D_{\triangle X}e\|_2 \\
&\leqslant 2^{-\frac{3}{2}}\sqrt{n}\ \|w_1+w_2\|_2^2 \\
&= 2^{-\frac{3}{2}}\sqrt{n}\ \left\|(D_XD_V)^{-\frac{1}{2}}\left(\tau\mu e - D_{\triangle X^a}D_{\triangle V^a}e - D_XD_Ve\right)\right\|_2^2 \\
&\leqslant 2^{-\frac{3}{2}}\sqrt{n}\left(\left\|(D_XD_V)^{-\frac{1}{2}}\left(\tau\mu e - D_XD_Ve\right)\right\|_2\right. \\
&\qquad \left. + \left\|(D_XD_V)^{-\frac{1}{2}}D_{\triangle X^a}D_{\triangle V^a}e\right\|_2\right)^2. \qquad (9)
\end{aligned}$$

Lets start by estimate

$$\left\|(D_XD_V)^{-\frac{1}{2}}\left(\tau\mu e - D_XD_Ve\right)\right\|_2.$$

Since $(x,u,v)\in\mathbb{V}_1(\theta)$, then

$$\|D_XD_Ve-\tau\mu e\|_2 \leqslant \|D_XD_Ve-\tau\mu e\|_1 \leqslant \theta\tau\mu$$

and $\min\limits_i\{x_iv_i\}\geqslant(1-\theta)\tau\mu$, therefore, one gets

$$\begin{aligned}
\left\|(D_XD_V)^{-\frac{1}{2}}\left(\tau\mu e - D_XD_Ve\right)\right\|_2 &\leqslant \left(\sum_{i\in\mathbb{I}}\frac{1}{x_iv_i}\right)^{\frac{1}{2}}\|D_XD_Ve-\tau\mu e\|_2 \\
&\leqslant \frac{\sqrt{n}\theta\tau\mu}{\sqrt{(1-\theta)\tau\mu}}. \qquad (10)
\end{aligned}$$

To estimate

$$\left\|(D_XD_V)^{-\frac{1}{2}}D_{\triangle X^a}D_{\triangle V^a}e\right\|_2$$

one uses Lemma 1 and $\min\limits_i\{x_iv_i\}\geqslant(1-\theta)\tau\mu$, and obtains

$$\begin{aligned}
\left\|(D_XD_V)^{-\frac{1}{2}}D_{\triangle X^a}D_{\triangle V^a}e\right\|_2 &= \left(\sum_{i\in\mathbb{I}}\frac{(\triangle x_i^a\triangle v_i^a)^2}{x_iv_i}\right)^{\frac{1}{2}} \\
&\leqslant \frac{\sqrt{n}\theta^2\tau\mu}{4(1-\theta)\sqrt{(1-\theta)\tau\mu}}. \qquad (11)
\end{aligned}$$

Applying (10) and (11) in (9), one gets

$$\|D_{\triangle V} D_{\triangle X} e\|_1 \leqslant 2^{-\frac{11}{2}} n^{\frac{3}{2}} \frac{\left(4\theta - 3\theta^2\right)^2}{(1-\theta)^3} \tau\mu.$$

Since $0 \leqslant \theta < 1$, the function $g(\theta) = \left(4\theta - 3\theta^2\right)^2$ has a maximum at $\theta = \frac{2}{3}$, then

$$0 \leqslant g(\theta) = \left(4\theta - 3\theta^2\right)^2 \leqslant \frac{16}{9}.$$

Therefore,

$$\|D_{\triangle V} D_{\triangle X} e\|_1 \leqslant n^{\frac{3}{2}} \frac{1}{18\sqrt{2}(1-\theta)^3} \tau\mu.$$

To obtain the relation between the duality measure of the next iterate of *Algorithm BP* with the current one the next result, proved in [18], will be used.

Lemma 3. *Consider μ defined in (2). Then,*

$$\mu^+ = \Big(1 - \delta(\lambda)\Big)\mu$$

where

$$\delta(\lambda) = \lambda \left(1 - \frac{n}{\beta(n)} + \frac{\triangle v^{aT} \triangle x^a}{v^T x}\right).$$

The previous results allows to obtain an upper bound for the steplength taken by *Algorithm BP*, as shown in the following theorem.

Theorem 1. *Consider that the current iterate of Algorithm BP, (x, u, v), is in $\mathbb{V}_1(\theta)$. Suppose that $(\triangle x^a, \triangle u^a, \triangle v^a)$ is the solution of (3) and that $(\triangle x, \triangle u, \triangle v)$ is the solution of (4). Then, the maximum step size, λ, that keeps the next iterate of Algorithm BP, (x^+, u^+, v^+), in $\mathbb{V}_1(\theta)$, satisfies*

$$\lambda \leqslant \frac{9\sqrt{2}(4\theta - 3\theta^2)(1-\theta)^2}{2n^{\frac{3}{2}}}, \qquad k \geqslant 0, \quad n \geqslant 5.$$

Proof. The aim, in this proof, is to find the maximum steplength $\lambda \geqslant 0$ that guarantees $(x^+, u^+, v^+) \in \mathbb{V}_1(\theta)$.

As $\triangle x^{aT} \triangle v^a \geqslant 0$, it is assured that $\mathbb{I}_+ \neq \emptyset$. Thus, consider $i \in \mathbb{I}_+$. Then,

$$x_i^+ v_i^+ = (1-\lambda) x_i v_i + \lambda\tau\mu - \lambda \triangle x_i^a \triangle v_i^a + \lambda^2 \triangle x_i \triangle v_i.$$

To satisfy the condition $(x^+, u^+, v^+) \in \mathbb{V}_1(\theta)$, lets consider $x_i^+ v_i^+ - \tau\mu^+$ and Lemma 3. Therefore, one gets

$$x_i^+ v_i^+ - \tau\mu^+ = (x_i v_i - \tau\mu)\Bigg(1 - \lambda\Bigg(1 + \frac{\triangle x_i^a \triangle v_i^a}{x_i v_i - \tau\mu} - \lambda\frac{\triangle x_i \triangle v_i}{x_i v_i - \tau\mu}$$

$$-\frac{\tau\mu}{x_i v_i - \tau\mu}\left(1 + \frac{\triangle v^{aT}\triangle x^a}{v^T x} - \frac{n}{\beta(n)}\right)\Bigg)\Bigg).$$

Hence, in order to obtain the maximum λ, such that $(x^+, u^+, v^+) \in \mathbb{V}_1(\theta)$, consider

$$\frac{\triangle x_i^a \triangle v_i^a}{x_i v_i - \tau\mu} - \lambda\frac{\triangle x_i \triangle v_i}{x_i v_i - \tau\mu} - \frac{\tau\mu}{x_i v_i - \tau\mu}\left(1 + \frac{\triangle v^{aT}\triangle x^a}{v^T x} - \frac{n}{\beta(n)}\right) \geqslant$$

$$\geqslant \frac{\triangle v^{aT}\triangle x^a}{v^{kT} x^k} - \frac{n}{\beta(n)},$$

consequently, obtaining

$$\triangle x_i^a \triangle v_i^a + (x_i v_i)\left(-\frac{\triangle v^{aT}\triangle x^a}{v^T x} + \frac{n}{\beta(n)}\right) - \tau\mu \geqslant \lambda\triangle x_i \triangle v_i,$$

which allows to get

$$\triangle x_i^a \triangle v_i^a + x_i v_i - \tau\mu \geqslant \lambda\triangle x_i \triangle v_i. \tag{12}$$

Now, using Lemma 1, one has

$$\triangle x_i^a \triangle v_i^a + x_i v_i - \tau\mu \leqslant \frac{1}{4}\frac{\theta^2}{(1-\theta)}\tau\mu + \theta\tau\mu$$

and, using Proposition 1, one can obtain

$$\triangle x_i^a \triangle v_i^a + x_i v_i - \tau\mu \geqslant \lambda\frac{n^{\frac{3}{2}}}{18\sqrt{2}(1-\theta)^3}\tau\mu \geqslant \lambda\triangle x_i \triangle v_i.$$

Therefore,

$$\lambda \leqslant \frac{9\sqrt{2}(4\theta - 3\theta^2)(1-\theta)^2}{2n^{\frac{3}{2}}}.$$

In order to prove Theorem 2, which indicates an upper bound for the number of iterations in which *Algorithm BP* stops and allows to conclude that it has polynomial complexity, Theorem 3.2 of [20, p. 61] (here transcribed and renamed as Lemma 4) is required.

Lemma 4. *Let $\epsilon \in]0,1[$ be given. Suppose that our algorithm for solving the Karush-Kuhn-Tucker conditions associated with the standard primal-dual pair of Linear Programming problems generates a sequence of iterates that satisfies*

$$\mu^+ \leqslant \left(1 - \frac{\delta}{n^\omega}\right)\mu$$

for some positive constants δ and ω. Suppose too that the starting point (x^0, u^0, v^0) satisfies

$$\mu^0 \leqslant \frac{1}{\epsilon^\kappa}$$

for some positive constant κ. Then there exists an index K with

$$K = O\left(n^\omega \left|\log(\epsilon)\right|\right)$$

such that

$$\mu \leqslant \epsilon \quad \text{for all } \kappa \geqslant K.$$

An upper bound for the number of iterations in which *Algorithm BP* stops is next established.

Theorem 2. *Let $\epsilon \in]0,1[$ and $\triangle x^{a^T} \triangle v^a \geqslant 0$. Algorithm BP stops after at most $O(n^{\frac{3}{2}} |\log(\epsilon)|)$ iterations with a solution for which $x^T v \leqslant \epsilon$.*

Proof. Consider μ define as in (2) and Lemma 3. In order to obtain an upper bound for μ^+ a lower bound for $\delta(\lambda)$ must be found. Since $\triangle x^{a^T} \triangle v^a \geqslant 0$, and $n\sqrt{n} \leqslant \beta(n) \leqslant n^2$, then

$$\delta(\lambda) \geqslant \lambda\left(1 - \frac{1}{\sqrt{n}}\right).$$

Using Theorem 1, for $n \geqslant 5$ and

$$\lambda = \frac{9\sqrt{2}(4\theta - 3\theta^2)(1-\theta)^2}{2n^{\frac{3}{2}}},$$

one gets

$$\mu^+ \leqslant \left(1 - \lambda\left(1 - \frac{1}{\sqrt{n}}\right)\right)\mu \leqslant \left(1 - \frac{9\sqrt{2}(4\theta - 3\theta^2)(1-\theta)^2}{4n^{\frac{3}{2}}}\right)\mu.$$

Finally, applying Lemma 4 the proof is complete.

This last result allows to conclude that this feasible version of the algorithm developed by Bastos and Paixão that uses the 1-norm neighborhood, *Algorithm BP*, is Q-linear convergent.

4 Conclusions

The analysis of the performance of interior-point algorithms continues to rely predominantly on their numerical effectiveness, since, in practice, worst-case complexity bounds tend to be overly conservative and not very representative of the behavior observed in real-world instances. Nevertheless, the rigorous study of asymptotic complexity remains a central element in theoretical research in this field. Several authors have demonstrated that the existence of a guarantee that the number of iterations grows in a controlled manner with the problem size is not merely an academic result, but a relevant indicator of a method's structural robustness [2,7,10–12,14,16,17,19,21–25]. An algorithm that satisfies such a guarantee offers greater predictability and stability in the face of adverse instances, reducing the probability of severe performance degradation. Thus, although practical efficiency is an important criterion when comparing methods performances, theoretical bounds that ensure polynomial computational effort constitutes an important complement that reinforces confidence in the algorithm's overall reliability and its ability to maintain controlled behavior even in extreme scenarios.

In this work, it is proved that, at each iteration of the *Algorithm BP*, the step size computed is bounded by $\frac{9\sqrt{2}(4\theta-3\theta^2)(1-\theta)^2}{2n^{\frac{3}{2}}}$, where $\theta \in [0,1[$, therefore, obtaining an $O\left(n^{\frac{3}{2}}|\log(\epsilon)|\right)$ iteration complexity bound; thus, proving that the feasible version of the 1-norm neighborhood algorithm developed by Bastos and Paixão in [3] has polynomial iteration complexity and is Q-linear convergent. In this sequel, it can be concluded that the obtained results show that it could be worth to implement the proposed variant of the Bastos and Paixão algorithm and analyze the results of its application to transportation and assignment class of problems.

Finally, it is important to observe that in this paper the authors concentrate on the theoretical aspects of the 1-norm neighborhood, specifically focusing on the convergence and complexity analysis of the algorithm under analysis. However, although no computational results have been included, the authors recognize that numerical evaluation of the algorithm is important to illustrate the practical impact of different geometric choices in the neighborhood of the central path. Thus, as future work, they intend to further study the algorithm, not only to improve its convergence results, but also to proceed with its implementation, conducting numerical experiments to compare the 1-norm neighborhood approach with the ones that use classical neighborhoods.

Acknowledgements. The authors were partially financed at Centro de Matemática da Universidade do Minho by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Project UID/00013/2025. https://doi.org/10.54499/UID/00013/2025

References

1. Al-Jeiroudi, G., Gondzio, J.: Convergence analysis of the inexact infeasible interior-point method for linear optimization. J. Optim. Theory Appl. **141**, 231–247 (2009)
2. Almeida, R., Teixeira, A.: On the convergence of a predictor-corrector variant algorithm. TOP **23**(2), 401–418 (2015)
3. Bastos, F., Paixão, J.: Interior-point approaches to the transportation and assignment problems on microcomputers. Investigação Operacional **13**(1), 3–15 (1993)
4. Benhadid, A., Merahi, F.: Complexity analysis of an interior-point algorithm for linear optimization based on a new parametric kernel function with a double barrier term. Numer. Algebra, Control Optim. **13**(2), 224–238 (2023)
5. Dantzig, G.B.: Linear programming, in problems for the numerical analysis of the future. In: Proceedings of the Symposium on Modern Calculating Machinery and Numerical Methods, Applied Mathematics Series, vol. 15, pp. 18–21 (1951)
6. Dexter, G., Chowdhury, A., Avron, H., Drineas, P.: On the convergence of inexact predictor-corrector methods for linear programming. In: International Conference on Machine Learning, pp. 5007–5038 (2022)
7. Güler, O., Ye, Y.: Convergence behavior of interior-point algorithms. Math. Program. **60**, 215–228 (1993)
8. Liu, L., Hua, T.: A polynomial interior-point algorithm with improved iteration bounds for linear optimization. Jpn. J. Ind. Appl. Math. **41**(2), 739–756 (2024)
9. Karmarkar, N.K.: A new polinomial-time algorithm for linear programming. Combinatorica **4**, 373–395 (1984)
10. Kheirfam, B.: A new predictor-corrector infeasible interior-point algorithm for linear optimization in a wide neighborhood. Fund. Inform. **177**, 141–156 (2020)
11. Kheirfam, B., Nasrollahi, A.: A wide neighborhood predictor-infeasible corrector interior-point algorithm for linear optimization. Optim. Lett. **14**(8), 2549–2563 (2020). https://doi.org/10.1007/s11590-020-01573-4
12. Korzak, J.: Convergence analysis of inexact infeasible-interior-point algorithms for solving linear programming problems. SIAM J. Optim. **11**(1), 133–148 (2000)
13. Mehrotra, S.: On the implementation of a primal-dual interior point method. SIAM J. Optim. **2**, 575–601 (1992)
14. Mizuno, S.: Polynomiality of infeasible interior-point algorithms for linear programming. Math. Program. **67**, 109–119 (1994)
15. Nesterov, Y., Nemirovski, A.: Interior-Point Polynomial Algorithms in Convex Programming. SIAM, Philadelphia, USA (1994)
16. Potra, F.: A quadratically convergent predictor-corrector method for solving linear programs from infeasible starting points. Math. Program. **67**, 383–406 (1994)
17. Salahi, M., Peng, J., Terlaky, T.: On Mehrotra-type predictor-corrector algorithms. SIAM J. Optim. **18**(4), 1377–1397 (2007)
18. Teixeira, A., Almeida, R.: On the complexity of a linear programming predictor-corrector algorithm using the two norm neighborhood. In: International Conference on Optimization, Learning Algorithms and Applications. Communications in Computer and Information Science, vol. 1754. pp. 533–545 (2022)
19. Yang, X., Yin, H.: A Mizuno-Todd-Ye predictor-corrector infeasible-interior-point method with the l1-norm wide neighborhood for linear programming. Appl. Math. Sci. **16**, 511–528 (2022)
20. Wright, S.J.: Primal-Dual Interior-Point Methods. SIAM, Philadelphia, USA (1997)

21. Wright, S., Zhang, Y.: A superquadratic infeasible interior point method for linear complementarity problems. Math. Program. **73**, 269–289 (1996)
22. Zhang, Y., Zhang, D.: On polynomiality of the Mehrotra-type predictor-corrector interior-point algorithms. Math. Program. **68**, 303–318 (1995)
23. Zhang, Y.: On the convergence of a class of infeasible interior-point methods for the horizontal linear complementarity problem. SIAM J. Optim. **4**(1), 208–227 (1994)
24. Zhang, Z., Zhang, Y.: A Mehrotra-type predictor-corrector algorithm with polynomiality and Q-subquadratic convergence. Ann. Oper. Res. **62**, 131–150 (1996)
25. Zhang, Y., Tapia, R.A., Dennis, J.E.: On the superlinear and quadratic convergence of primal-dual interior point linear programming algorithms. SIAM J. Optim. **2**, 304–324 (1992)

Mixed-Integer Linear Programming for Production Scheduling with Robot Allocation in a Textile Company

Maria Coelho Lima[1(✉)], Carina Pimentel[1,2], and Filipe Alvelos[1,2]

[1] ALGORITMI Research Center/LASI, School of Engineering, University of Minho, Campus of Azurém, 4800-058 Guimarães, Portugal
maria.afonso.coelho.lima@gmail.com

[2] Department of Production and Systems, School of Engineering, University of Minho, Campus of Azurém, 4800-058 Guimarães, Portugal

Abstract. This article addresses production scheduling challenges in the luxury textile industry, with a particular focus on robot-tufted carpet manufacturing. The study was developed in collaboration with a Portuguese consultancy company and targets the complexity arising from high product customization, strict delivery deadlines, and limited production resources.

A production scheduling model based on Mixed-Integer Linear Programming (MILP) is proposed to optimize the allocation and sequencing of carpets in the main robot-assisted production task. The model incorporates key operational constraints, including robot compatibility, yarn availability, and production deadlines. The objective is to minimize critical delays of the task related to completion within two days of the delivery date, promote completion at least eight days in advance, minimize makespan per robot production type, and enhance production anticipation while maintaining schedule feasibility. To guarantee computational efficiency in an industrial setting, solver parameter tuning strategies were implemented. To further improve performance, the Fix-and-Optimize heuristic was tested; however, its impact was limited for the problem scale considered. Consequently, a warm-start strategy based on existing production plans was adopted to provide high-quality initial solutions and enhance solver convergence.

The proposed approach was validated using real industrial data and benchmarked against current scheduling practices. Results show that optimized schedules increased the proportion of orders completed at least two days before delivery by 1.9% and at least eight days before by 3.7%, while also reducing makespan.

Overall, the study demonstrates the effectiveness of mathematical optimization as a decision-support tool for complex production environments, reducing scheduling time while improving service levels and operational performance in the luxury manufacturing sector.

Keywords: Textile Industry · Production Scheduling · Mixed-Integer Linear Programming

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 314–331, 2027.
https://doi.org/10.1007/978-3-032-30521-3_20

1 Introduction

The textile sector is dynamic and subject to constant market fluctuations, making it essential to adopt advanced planning techniques that improve both responsiveness and decision quality [1]. At the same time, increasing product customization raises production variability and complexity in planning and forecasting. This scenario calls for the integration of advanced data-driven approaches [2].

This work is framed within a Portuguese textile company dedicated to the production of luxury carpets, operating two manufacturing units. The company follows a made-to-order strategy, producing both catalog and bespoke products. Approximately 70% of orders correspond to customizable catalog models, while the remaining 30% are fully bespoke, requiring flexible and responsive production planning.

The company relies on four main production techniques: flatweave, hand-knotted, hand-woven, and robot-tufted. Among these, robot-tufted is the most technologically advanced and operationally significant, representing the majority of current output. This process relies on automated robotic systems to inject yarn into a stretched carpet base according to predefined digital trajectories.

The robot-tufted process involves multiple departments, including order processing, design, yarn availability, and scheduling. Currently, scheduling is manual and relies on planners' experience, often leading to inefficiencies and an average delay rate of around 12%. These challenges motivated the development of an optimization-based decision-support model capable of generating sequenced production plans based on the order portfolio and operational constraints. In particular, this work proposes a Mixed-Integer Linear Programming model that integrates allocation and sequencing decisions while incorporating key industrial constraints, and evaluates its performance using real production data.

The remainder of this paper is organized as follows. Section 2 presents background on production scheduling. Section 3 introduces the proposed model. Section 4 describes the solution approach. Section 5 presents the results, and Sect. 6 concludes the paper.

2 Theoretical Background

Production scheduling is a decision-making process used in manufacturing and service sectors to allocate resources such as labor, machines, or capital to tasks in a defined sequence and over time [3].

Historically, scheduling was considered a simple step following production planning. However, modern manufacturing environments, characterized by high variability, customization, and complex resource interactions, require more advanced scheduling methodologies. Advances in decision support optimization make it increasingly feasible to replace manual spreadsheet-based planning with robust analytical tools capable of generating highly efficient solutions [4]. These methods help identify near-optimal solutions that minimize production time and improve resource utilization [5].

This section presents an overview of production scheduling, including problem classification, solution approaches, and computational efficiency strategies, followed by relevant applications in manufacturing contexts.

2.1 Classification of Scheduling Problems

Scheduling problems are commonly described using the three-field notation $\alpha|\beta|\gamma$ introduced in [7].

The first element, α, defines the machine environment, influencing problem complexity and enabling comparison between scheduling approaches. Common environments include single-machine, parallel-machine, and multi-machine shop systems.

Single-machine environments process tasks sequentially on one resource, while parallel-machine environments allow tasks to be processed by multiple machines, which may be identical, uniform, or unrelated. Multi-machine shop environments include flow shops, job shops, and open shops, each characterized by different routing and sequencing structures.

The second field, β, describes attributes and constraints associated with jobs, such as precedence relations, release dates, setup times, operation limits, and unit processing time assumptions. These features influence model complexity and solution choice.

The third field, γ, defines the performance objective used to evaluate a schedule, such as makespan, lateness, tardiness, earliness, setup time, and weighted objectives reflecting job priorities.

2.2 Scheduling Approaches

Since the mid-1950s, scheduling problems have increased in number and complexity [8]. This has led to a range of approaches, from simple heuristics to advanced mathematical models.

A common approach uses priority-based dispatching rules to select the next task for processing. Dispatching rules have a significant role within the dynamic context due to their ease of implementation and compatibility with the dynamic nature of manufacturing systems [9]. Typical examples include First-Come, First-Served, Earliest Due Date, Shortest or Longest Processing Time, and the Minimum Slack rule.

Mathematical programming, particularly Mixed-Integer Linear Programming, has become one of the most widely explored methods for process scheduling due to its rigor, flexibility, and strong modeling capabilities [10]. Scheduling models can use discrete or continuous time, with continuous formulations often preferred due to lower approximation errors and computational burden.

Continuous-time models can be further categorized as immediate precedence or general precedence formulations. Immediate precedence models enforce that one task starts directly after another on the same machine [11], whereas general precedence formulations impose only relative ordering constraints between

tasks [12]. A comparative study found that the general precedence formulation is the most computationally efficient among the commonly used models [13], motivating its adoption in this research work.

Constraint Programming (CP) is a declarative paradigm for solving combinatorial problems by explicitly defining constraints that feasible solutions must satisfy [14]. These approaches are valued for their modeling flexibility and integration in industrial decision-support systems.

Finally, Metaheuristics balance exploration and exploitation to obtain high-quality solutions efficiently [15]. Common methods include Simulated Annealing, Genetic Algorithms, and Tabu Search, which are effective for complex problems despite not guaranteeing optimality.

2.3 Computational Efficiency and Applications

Scheduling problems are computationally complex, making efficiency critical in industrial applications. Performance can be improved through model simplification, decomposition, warm starts, and solver tuning.

Model simplification reduces problem size by relaxing non-critical constraints or aggregating variables. Decomposition methods divide large problems into smaller subproblems, for example, Fix-and-Optimize iteratively fixes subsets of variables and re-optimizes the remaining ones [16]. Warm-start techniques provide an initial feasible solution to guide the branch-and-bound search process. Solver parameter tuning can also significantly affect performance [17].

Optimization-based approaches have been applied to textile and related manufacturing industries. A MILP model for production planning in a luxury textile company was presented in [18], incorporating machine capacity and material constraints. A weighted-tardiness problem in a textile manufacturing context was addressed using a hybrid MILP and metaheuristic approach in [19]. Earlier work described scheduling challenges in carpet production within traditional non-automated environments [20].

While prior studies highlight the relevance of mathematical programming in textile manufacturing, limited work addresses automated carpet production with jointly coordinated robotic resources, material availability, and delivery constraints. This work fills that gap with an integrated MILP model for robot-assisted production.

3 Optimization Modeling

The robot-tufted production stage begins with robotic yarn injection and is followed by finishing operations, which include trimming, and quality assurance steps. Since the first stage determines the start of the subsequent processes, the scheduling model developed in this work focuses specifically on this stage.

The operational rules governing the injection stage were first mapped and then formalized as a MILP model. In this formulation, products are assigned and sequenced on compatible robots according to their typology (injection type).

Structure allocation depends solely on dimensional compatibility, with a mandatory 20 cm safety margin. Production can start only two days after the latest yarn arrival and must respect the working calendar, labor capacity, planned stoppages, and setup requirements. The first two days of the schedule are kept fixed.

Special cases, such as carpets that are too small or too large to follow the standard rules, introduce additional timing and assignment constraints. No explicit planning horizon is imposed, as all pending orders must be scheduled.

Based on these constraints and planning needs, the model was designed with the following objectives:

- Minimize the number of delays, where a minor delay is defined as an injection stage ending less than eight days before the delivery date, reducing the buffer time for finishing and shipping, and a major delay occurs when injection ends less than two days before the delivery date, making it almost impossible to meet the delivery deadline;
- Minimize the time each robot typology takes to complete all assigned orders;
- Minimize the time lag between material availability and the initiation of production.

While the first two objectives were included in the initial formulation, the third was introduced after preliminary tests revealed idle gaps despite material availability.

3.1 Data Preparation

To simplify the formulation and reduce computational complexity, several preparatory steps were applied before modeling. Instead of incorporating all conditions directly as constraints, part of the filtering and structuring was conducted in advance through data preparation.

The first filtering step selected unplanned products that had already undergone color analysis, ensuring yarn requirements and availability were known.

Robot information was then collected, including typology, dimensional capacity, manufacturing unit, and activity status, retaining only active robots. A similar procedure was applied to structures, considering size, activity status, and manufacturing unit. The company's manufacturing calendar was also processed to obtain the start and end times of planned robot stoppages.

Product data were consolidated from multiple sources, starting with the verification of yarn availability. For products without yarn in stock or a defined arrival date, an artificial arrival time was assigned by the company. This was set to five days from the current date if the delivery date was within twenty-one days, and to twelve days otherwise. This approach allowed these products to be scheduled while maintaining a realistic buffer for yarn availability. For each product, the latest arrival among all required yarns was selected, and two additional days were added to account for yarn preparation.

Products were also assigned a typology based on the robot previously allocated by planners. Although this assignment was necessary, future work should define typology independently to avoid reliance on initial plans. Production duration was then calculated by dividing the effective area by the injection rate estimated by planners, ensuring consistency with company practice.

A key methodological decision was the adoption of continuous time rather than indexed time periods. This choice increases the realism and flexibility of the formulation by allowing tasks to be scheduled with greater precision. This choice is supported by the characteristics of the robot-tufted process. Robots operate autonomously during production and require human intervention only during setup operations. As a result, production can continue uninterrupted even during worker breaks. All planned machine stoppages are known in advance and incorporated into the model as scheduled interruptions. All relevant dates (delivery deadlines, yarn arrivals, and stoppages) were converted into hours measured from the start of the planning horizon (hour zero).

Building on this preprocessing, several datasets were created to encode feasibility conditions, ensuring that only feasible allocations and sequencing possibilities were considered by the model. These datasets form the basis of the mathematical formulation and are presented next.

- PR: all feasible product–robot combinations, filtered by compatibility of typology and dimensions, as well as manufacturing unit restrictions. For Manufacturing Unit 2, only products with all yarns already in stock were included.
- PS: all feasible product–structure allocations, filtered by dimensions with a mandatory security margin of 20 cm on all borders, along with the same manufacturing unit restrictions.
- C: the set of all carpet samples, which must be assigned to a robot but not to a structure, since they are handled differently in production.
- PM: the set of product–manufacturing unit combinations, created only for products with yarn in stock, which can potentially be produced in either unit.
- PP' and $PP'R$: pairs of products with the same typology, and their feasible allocations to the same robot, respectively, supporting the sequencing constraints. During dataset creation, setup times between consecutive products were specified: fifteen minutes for standard machine preparation, reduced to five minutes for products from the same order since the same yarn is used. This differentiation encourages the model to schedule products from the same order consecutively on the same robot.
- F: products with fixed allocations for the first planning days, which the model cannot modify. The number of fixed days can be adjusted as required.
- PKR and PK: respectively, combinations of product–robot pairs (from PR) with stoppages, used to model robot-dependent interruptions, and combinations of products with stoppages, used for robot-independent interruptions.
- Special carpets are subdivided into three datasets:
  - SD: descending carpets are carpets too large to be produced in a single pass and must be partially produced, repositioned (descended), and then completed. In the data, they appear as two products with different map

codes but the same product code, and a dataset was created to link these pairs so the model treats them correctly.
- *SS*: spliced carpets are also large carpets but produced in two parts and later joined. They appear as separate products with different map and product codes, making them impossible to identify automatically. A dedicated dataset was therefore used to detect these pairs, which were then consolidated into a simpler dataset for the model.
- *SSS*: splice cases where the first half was previously produced and the second must now be scheduled, typically as rework. This dataset retrieves the splice information and the robot used for the first half, ensuring the second half is produced on the same robot.

In addition, carpets small enough to be grouped were treated as a single aggregated product with combined dimensions, since they are injected together in the same operation, without requiring a dedicated dataset.

In summary, the data preparation phase ensured model feasibility while keeping the formulation manageable by providing structured and accurate inputs.

3.2 Mathematical Formulation

The formulation is structured into indices and sets, parameters, decision variables, constraints, and an objective function.

Indices and Sets. Indices and sets provide the structural basis of the formulation: indices reference individual elements, while sets define the collection of feasible elements.

For this model, the basic indices and their respective sets are:

- $p \in P$: products
- $r \in R$: robots
- $t \in T$: injection typologies
- $s \in S$: structures
- $m \in M$: manufacturing units
- $k \in K$: planned stoppages

Additional feasibility datasets such as PR, PS, C, PM, PP', $PP'R$, F, PKR, PK, SD, SS, and SSS are incorporated into the formulation.

Parameters. Parameters define the input characteristics of the scheduling problem, describing the attributes of products, resources, and stoppages. All time-related parameters were expressed in hours for consistency.

- $Duration_p$: production time for product p
- $Typology_p, Typology_r$: typology of product p and robot r
- $Delivery_p$: delivery deadline of product p

- $Yarn_p$: time of the latest yarn arrival for product p
- $Setup_{pp'}$: setup time between products p and p', depending on order relation
- $KStart_k, SEnd_k$: start and end times of stoppages k
- $KDuration_k$: duration of stoppage k
- $PlannedStart_p$: start time for product p in previous plan

Decision Variables. Decision variables represent the model's choices for building a feasible schedule, defining product allocation, sequencing, and operation timing. Both binary and continuous variables are used, with time expressed in hours.

$$X_{pr} = \begin{cases} 1 & \text{if product } p \text{ is assigned to robot } r \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$Y_{ps} = \begin{cases} 1 & \text{if product } p \text{ is assigned to structure } s \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$Z_{pp'} = \begin{cases} 1 & \text{if product } p \text{ precedes } p' \text{ on the same robot} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$W_{p,k} = \begin{cases} 1 & \text{if stoppage } k \text{ interrupts product } p \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$B_{p,k} = \begin{cases} 1 & \text{if product } p \text{ finishes before stoppage } k \text{ begins} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$Start_p$: start time of product production p.

End_p: end time of product production p.

$Makespan_t$: completion time of the last product for each injection typology t.

$Tardiness2_p$: delay recorded when the production of product p ends less than two days before its delivery date.

$Tardiness8_p$: delay recorded when the production of product p ends less than eight days before its delivery date.

$StartAfterYarn_p$: time between yarn availability and start of product p.

Constraints. Constraints define the operational and logical rules that any feasible solution must satisfy, ensuring the production plan respects resource, timing, and sequencing requirements.

Allocation constraints, presented below, ensure each product is assigned to available and appropriate manufacturing resources.

$$\sum_{r \in R:(p,r) \in PR} X_{pr} = 1, \ \forall p \in P \tag{6}$$

$$\sum_{s \in S:(p,s) \in PS} Y_{ps} = 1, \ \forall p \in P \setminus C \tag{7}$$

$$\sum_{s \in S:(p,s) \in PS} Y_{ps} = 0, \ \forall p \in C \tag{8}$$

$$\sum_{r \in R:(p,r) \in PR, \ M(r)=m} X_{pr} = \sum_{s \in S:(p,s) \in PS, \ M(s)=m} Y_{ps}, \forall (p,m) \in PM \setminus C \tag{9}$$

First, each product must be assigned to exactly one robot, as defined in Constraint (6). Then, the products must also be assigned to exactly one structure, as specified in Constraint (7), with the exception of sample products, which cannot be allocated to any structure (Constraint (8)). Furthermore, each product must be assigned to both a robot and a structure within the same manufacturing unit, as enforced by Constraint (9). Since sample products are not assigned to any structure, this constraint does not apply to them.

Following this, production constraints ensure the temporal feasibility of the schedule, taking into account material availability and production duration.

$$Start_p \geq Yarn_p, \ \forall p \in P \tag{10}$$

$$End_p \geq Start_p + Duration_p + \sum_{k:(p,k) \in PK} (W_{p,k} \times KDuration_k), \ \forall p \in P \tag{11}$$

Initially, the production of each carpet can only start once the required yarn has arrived, as defined in Constraint (10). Additionally, the completion time of each carpet must be at least equal to its start time plus the production duration, including possible interruptions due to stops, as specified in Constraint (11).

Robot sequencing constraints ensure that two products assigned to the same robot ($X_{pr} = 1$ and $X_{p'r} = 1$) are produced sequentially without overlap, considering setup times. This is enforced through binary sequencing variables and a big-M constant, which exceeds any possible start or end time within the planning horizon.

$$Start_{p'} + M(1 - X_{pr}) + M(1 - X_{p'r}) + M(1 - Z_{pp'}) \geq End_p + Setup_{pp'}, \\ \forall (p, p', r) \in PP'R \tag{12}$$

$$Start_p + M(1 - X_{pr}) + M(1 - X_{p'r}) + MZ_{pp'} \geq End_{p'} + Setup_{p'p}, \\ \forall (p, p', r) \in PP'R \tag{13}$$

Constraint (12) ensures that if product p precedes product p' on the same robot, then the start of p' must occur after the end of p plus the corresponding setup time. Conversely, Constraint (13) enforces that if product p' precedes product p, then the start of p must not occur earlier than the end of p' plus its respective setup time.

Fixed production constraints preserve predetermined schedules for specific products during the first production days while allowing flexibility for the remaining schedule, ensuring that fixed assignments, start times, and sequencing are respected.

$$X_{pr} = 1,\ \forall (p,r) \in F \tag{14}$$

$$Start_p \geq PlannedStart_p,\ \forall r \in R,\ p = \text{first fixed product of } r \tag{15}$$

As a first step, products in the set of fixed assignments must be allocated to their predefined robot, as defined in Constraint (14). To account for products whose production starts before the beginning of the planning horizon and has not yet been completed, the first product scheduled on each robot must have a start time no earlier than its previously planned starting hours, as specified in Constraint (15).

$$Z_{pp'} \geq 1,\ \forall (p,p') \in PP',\ p,p' \in F:\ PlannedStart_p \leq PlannedStart_{p'} \tag{16}$$

$$Z_{pp'} = 0,\ \forall (p,p') \in PP',\ p,p' \in F:\ PlannedStart_p \geq PlannedStart_{p'} \tag{17}$$

The relative order of fixed productions must also be preserved. If a fixed production p is scheduled to start before another fixed production p', then the binary variable $Z_{pp'}$ is set to 1, as defined in Constraint (16). Conversely, if p is scheduled to start after p', then $Z_{pp'}$ must be equal to 0, as specified in Constraint (17).

$$Z_{pp'} + M(1 - X_{p'r}) \geq 1,\ \forall (p,p') \in PP',\ p \in F,\ p' \notin F \tag{18}$$

$$Z_{pp'} + M(1 - X_{p'r}) \leq 0,\ \forall (p,p') \in PP',\ p \notin F,\ p' \in F \tag{19}$$

Finally, to avoid conflicts between fixed and non-fixed productions assigned to the same robot ($X_{pr} = 1$), sequencing rules ensure that non-fixed productions are scheduled only after fixed ones. This requirement is enforced by Constraints (18) and (19).

Stoppage constraints ensure that products are scheduled correctly with respect to machine stoppages, considering whether they are interrupted, completed beforehand, or postponed. Binary variables $W_{p,par}$ and $B_{p,par}$ are used to model whether product p is interrupted by stoppage k or completed before it, respectively. For all these constraints, the product must be allocated to the robot to which the stoppage refers ($X_{pr} = 1$), enforced through the big-M approach.

$$Start_p \leq KStart_k + (1 - W_{p,k})M + (1 - X_{pr})M,\ \forall (p,k,r) \in PKR \tag{20}$$

$$\begin{aligned} &Start_p + Duration_p + \sum_{\substack{(p,k',r) \in PKR: \\ KStart_{k'} < KStart_k}} (W_{p,k'} \times KDuration_{k'}) \geq \\ &KStart_k - (1 - W_{p,k})M - (1 - X_{pr})M, \forall (p,k,r) \in \text{ PKR} \end{aligned} \tag{21}$$

Initially, for a product to be interrupted by a stoppage ($W_{p,k} = 1$), it must start before the stoppage begins, as defined in Constraint (20). Furthermore, its production must extend beyond the start time of the stoppage, taking into account its duration and any previous stoppages that may have occurred, since a product can be interrupted by more than one stoppage, as specified in Constraint (21).

$$\begin{aligned} &Start_p + Duration_p + \sum_{\substack{(p,k',r) \in PKR: \\ KStart_{k'} < KStart_k}} (W_{p,k'} \times KDuration_{k'}) \geq \\ &KStart_k + (1 - X_{pr})M + (1 - B_{p,k})M, \ \forall (p,k,r) \in \text{ PKR} \end{aligned} \tag{22}$$

If a product finishes before the stoppage begins, it is not interrupted. This case is modeled in Constraint (22), which activates when $B_{p,k} = 1$.

$$Start_p \geq KEnd_k - W_{p,k}M - (1 - X_{pr})M - B_{p,k}M, \ \forall (p,k,r) \in PKR \tag{23}$$

Similarly, if a product starts only after the stoppage has ended, it is not interrupted ($W_{p,k} = 0$) and is not completed before the stoppage ($B_{p,k} = 0$), as defined in Constraint (23).

$$\sum_{p \in P} W_{p,k} \leq 1, \ \forall k \in K \tag{24}$$

$$W_{p,k} + B_{p,k} \leq 1, \ \forall (p,k) \in PK \tag{25}$$

Lastly, to ensure consistency, at most one product can be interrupted by each stoppage, as defined in Constraint (24), and a product must either be interrupted by a stoppage or completed before it, but not both, as specified in Constraint (25).

Special carpet constraints ensure correct handling of carpets produced in two parts (descending or spliced), maintaining allocation consistency and temporal feasibility.

$$X_{p1r} = X_{p2r}, \forall (p1,p2) \in (SD \cup SS), \ \forall r \in R: \ (p1,r) \in PR, \ (p2,r) \in PR \tag{26}$$

$$Y_{p1s} = Y_{p2s}, \ \forall (p1,p2) \in SD, \ \forall s \in S: \ (p1,s) \in PS, \ (p2,s) \in PS \tag{27}$$

First, allocation consistency must be preserved. For descending carpets or carpets with splices, both parts must be assigned to the same robot, as defined in Constraint (26). Moreover, descending carpets must also be assigned to the same structure, as specified in Constraint (27).

$$Start_{p2} \geq Start_{p1} + Duration_{p1} + \sum_{k:\,(p1,k)\in PK} (W_{p1,k} \times KDuration_k) + 24, \quad \forall (p1, p2) \in SD \tag{28}$$

$$Start_{p2} \leq Start_{p1} + Duration_{p1} + \sum_{k:\,(p1,k)\in PK} (W_{p1,k} \times KDuration_k) + 32, \quad \forall (p1, p2) \in SD \tag{29}$$

In addition to allocation, temporal consistency must be ensured. For descending carpets, the second part must begin at least 24 h and at most 32 h after the first part is completed, as defined in Constraints (28) and (29).

$$Start_{p2} \geq Start_{p1} + Duration_{p1} + \textstyle\sum_{k:\,(p1,k)\in PK} (W_{p1,k} \times KDuration_k) + 0.0833, \quad \forall (p1, p2) \in SS \tag{30}$$

$$Start_{p2} \leq Start_{p1} + Duration_{p1} + \textstyle\sum_{k:\,(p1,k)\in PK} (W_{p1,k} \times KDuration_k) + 0.5, \quad \forall (p1, p2) \in SS \tag{31}$$

For splices, the second part must begin at least 5 min (0.0833 h) and at most 30 min (0.5 h) after the first part is completed, as defined in Constraints (30) and (31).

$$X_{pr} = 1, \ \forall (p, r) \in SSS \tag{32}$$

In addition, for singular splices where the first part has already been completed, the second part must be assigned to the same robot as the one previously used, as defined in Constraint (32).

Finally, the objective function uses auxiliary variables capturing tardiness, makespan, and non-productive gaps, defined by the following constraints.

$$Tardiness2_p \geq End_p - (Delivery_p - 48), \ \forall p \in P \tag{33}$$

$$Tardiness8_p \geq End_p - (Delivery_p - 192), \ \forall p \in P \tag{34}$$

To begin with, the tardiness variables capture whether the production of a carpet is completed close to its delivery deadline. Constraint (33) accounts for delays occurring within two days (48 h) before the delivery date, while Constraint (34) captures delays occurring within eight days (192 h).

$$Makespan_t \geq End_p,\ \forall p \in P \text{ such that } Typology_p = t,\ \forall t \in T \tag{35}$$

The makespan is defined for each typology $t \in T$, ensuring that it is no smaller than the completion time of any product p whose typology corresponds to t, as expressed in Constraint (35).

$$StartAfterYarn_p \geq Start_p - Yarn_p,\ \forall p \in P \tag{36}$$

As a final step, to measure the time between the yarn arrival and the production start, Constraint (36) is introduced.

Objective Function. The objective function aims to minimize the weighted sum of several decision variables that represent different dimensions of scheduling performance, as shown in Eq. (37). These include makespan, tardiness relative to the two- and eight-day delivery thresholds, and waiting time between yarn availability and production start. Together, they provide a comprehensive measure of the overall efficiency of the schedule.

$$\begin{aligned} \min \Big(& w_1 \times \textstyle\sum_{t \in T} Makespan + w_2 \times \sum_{p \in P} Tardiness2_p \\ & + w_3 \times \textstyle\sum_{p \in P} Tardiness8_p + w_4 \times \sum_{p \in P} StartAfterYarn_p \Big) \end{aligned} \tag{37}$$

Each component of the objective function is multiplied by a weight (w_1, w_2, w_3, w_4), that defines its relative importance. These weights can be adjusted by the company according to its operational priorities. Minimizing makespan shortens the production horizon, minimizing tardiness relative to the two- and eight-day thresholds improves delivery reliability, and minimizing the time between yarn availability and production start reduces idle periods and improves resource utilization, ultimately enhancing schedule reliability and responsiveness.

4 Model Solving

The optimization model was implemented in Python using the PuLP library with the Gurobi Optimizer (version 11.0.3), a solver known for efficiently handling large-scale MILP problems.

The model was implemented through Python functions that progressively built the formulation, improving transparency and maintainability, with logging and progress tracking used to monitor solver performance. Once all variables, constraints, and the objective function were defined, the model was solved using Gurobi. During testing, the model exhibited high complexity, with several hundred thousand variables and constraints, leading to substantial computational effort and longer solution times. As a result, specific strategies to improve computational efficiency had to be explored, in order to obtain feasible solutions within a reasonable time.

To address computational challenges, several strategies were explored. Despite Gurobi's efficiency, solving the model was challenging, with slow feasibility detection and gap reduction. The optimality gap measures the relative difference between the best feasible solution and the best known lower bound, indicating how close the solution is to optimality.

The first approach involved tuning Gurobi's solver parameters to improve feasibility detection and convergence. The configuration that yielded the best results is summarized in Table 1.

Table 1. Best-performing Gurobi parameter configuration

Parameter	Value	Definition
MIPFocus	1	Prioritizes finding feasible solutions over proving optimality.
Heuristics	1	Increases the use of heuristics to obtain initial solutions faster.
Cuts	2	Applies additional cuts to strengthen the formulation.
Threads	4	Parallel threads to speed up the generation of schedules.

This configuration improved feasibility detection, but the solver often stalled at a given optimality gap. To address this, a Fix-and-Optimize heuristic was tested, which fixes part of the decision variables while optimizing the others to reduce problem size. Although this approach helped identify feasible solutions more quickly, it often plateaued without further improving the optimality gap.

Given these limitations, a warm start strategy was developed and proved more effective. A temporary constraint (Constraint (38)) was introduced to enforce the allocation of the same robot previously assigned to each order, when it appeared in an earlier production plan.

$$X_{pr} = 1, \quad \forall p \in p_{\text{planned}},\ r = r(planned)_p \tag{38}$$

The model was first solved under these restrictions for about five minutes to obtain an initial feasible solution. The constraint was then removed and the model restarted using this solution as a warm start. The solver was run for one hour, achieving a single-digit optimality gap with a good runtime–quality trade-off. This approach produced an initial solution almost instantly and improved the gap frequently, unlike the Fix-and-Optimize approach, where improvements required several minutes.

5 Results, Discussion and Implementation

Several production plans were generated and compared with the company's plans using key performance metrics. Five plans were considered, each with a different initial planning date and order set, covering a four-week horizon corresponding to the company's standard planning window.

The plans were generated during validation sessions in which the planning team directly tested the model and compared it with the corresponding company plans. The objective function weights (37) were initially defined based on the company's priorities, resulting in $w_1 = 1$, $w_2 = 12$, $w_3 = 2$, and $w_4 = 0.75$. However, this configuration led to an excessive focus on minimizing Dd-2 tardiness at the expense of makespan, revealing the sensitivity of the model to weight calibration. The weights were therefore adjusted based on empirical evaluation to $w_1 = 2$, $w_2 = 6$, $w_3 = 2$, and $w_4 = 0.75$, accounting for the relative magnitude of the objective components and yielding more balanced trade-offs.

Performance was evaluated using the average percentage of products completed on time across the five plans, considering two thresholds: completion at least two days (Dd-2) and eight days (Dd-8) before the delivery date (Fig. 1).

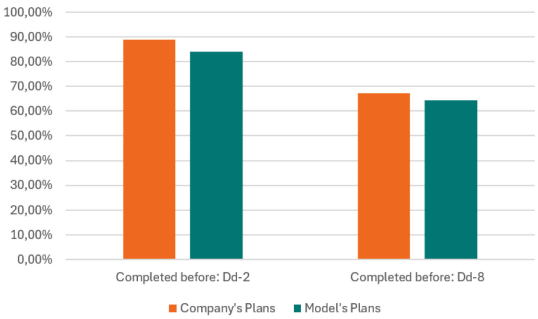


Fig. 1. Comparison of average tardiness (Dd-2 and Dd-8) between the company's plans and the model's plans

The model's plans initially underperformed compared with the company's plans in both thresholds, achieving 84.1% (Dd-2) and 64.3% (Dd-8), versus 88.9% and 67.1%, respectively. This gap was mainly driven by outdated yarn arrival data, which led to suboptimal scheduling decisions. In practice, planners often rely on informal updates regarding earlier yarn arrivals that are not reflected in the system, highlighting the importance of accurate yarn management and data updates.

To address this issue, yarn arrival dates were corrected based on actual availability, enabling a fairer comparison between the model and the company's plans.

As shown in Fig. 2, the yarn corrected model achieved 90.8% (Dd-2) and 70.8% (Dd-8) on-time completion, improving by 1.9 and 3.7% points over the company's plans.

These results demonstrate that scheduling performance is highly sensitive to input data reliability. However, when yarn genuinely arrives late, delivery performance cannot be recovered through scheduling alone, highlighting a limitation beyond the model's scope.

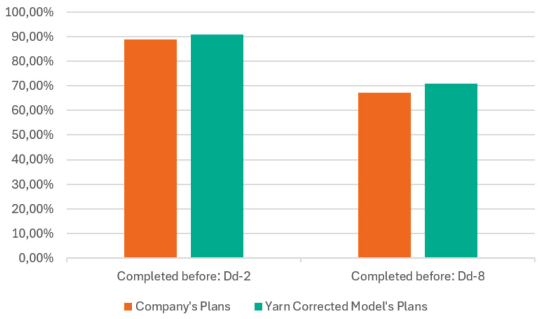


Fig. 2. Comparison of average tardiness (Dd-2 and Dd-8) between the company's plans and the yarn corrected model's plans

Finally, the makespan across different typologies was evaluated using the yarn-corrected model, as shown in Fig. 3.

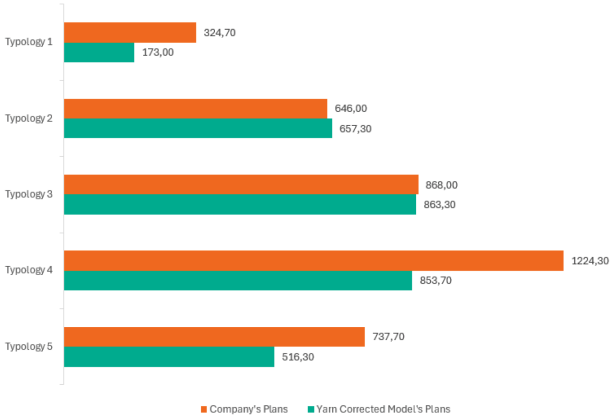


Fig. 3. Makespan comparison between the company's plans and the yarn corrected model's plans

The results show a reduction in makespan for most typologies compared with the company's plans, with the largest improvements observed in Typologies 1, 4, and 5, while Typologies 2 and 3 show comparable results. This indicates that the model is particularly effective for more time-intensive typologies, contributing to improved overall efficiency.

The yarn-corrected scenarios were used solely for evaluation purposes and are not directly implementable, as they rely on manually adjusted data. This reinforces the need for improved data integration and reliability to fully exploit the model's potential in practice.

Regarding implementation, the model has been integrated into the company's planning workflow. However, its current use is limited by the lack of structured product typology data, requiring manual assignment of compatible robots and preventing full automation. Input data are provided through an Excel interface connected to the company's database, after which the optimization application generates production plans for review and integration into the planning system.

6 Conclusion and Future Work

This work developed an optimization-based scheduling model for the robot-tufted production process in a luxury carpet manufacturing company. The model generates sequenced production plans while considering key operational constraints such as robot compatibility, yarn availability, setup times, stoppages, and delivery deadlines. The results show that, when reliable yarn availability data are available, the model can improve scheduling performance compared with the company's current manual planning approach, achieving higher on-time completion rates and reduced makespan in several production typologies.

Beyond its practical contribution, this work also addresses a research gap in the literature, as optimization-based scheduling for robot-tufted carpet production has not been extensively explored.

Future work should focus on advancing toward a fully automated scheduling process by integrating product typology directly from the design stage and improving the reliability of yarn availability data. Additionally, further research could explore improved strategies for handling special carpets, such as descending and spliced products, to further enhance scheduling efficiency.

Acknowledgments. This research was developed in collaboration with LTPlabs, a consulting company where the work was carried out and was supported by FCT – Fundação para a Ciência e Tecnologia under the R&D Unit Project Scope UID/00319/2025 - Centro ALGORITMI (ALGORITMI/UM), 10.54499/UID/00319/2025

References

1. Lorente-Leyva, L.L., Alemany, M.M.E., Peluffo-Ordóñez, D.H.: A conceptual framework for the operations planning of the textile supply chains: insights for sustainable and smart planning in uncertain and dynamic contexts. Comput. Ind. Eng. **187**, 109824 (2024). https://doi.org/10.1016/j.cie.2023.109824
2. Ray, S.: A quick review of machine learning algorithms. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 35–39 (2019). https://doi.org/10.1109/COMITCon.2019.8862451
3. Pinedo, M.L.: Scheduling: Theory, Algorithms, and Systems, 6th edn. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05921-6
4. Vieira, M., Pinto-Varela, T., Barbosa-Póvoa, A.P.: A model-based decision support framework for the optimisation of production planning in the biopharmaceutical industry. Comput. Ind. Eng. **129**, 354–367 (2019). https://doi.org/10.1016/j.cie.2019.01.045

5. Berthier, A., Yalaoui, A., Chehade, H., Yalaoui, F., Amodeo, L., Bouillot, C.: Unrelated parallel machines scheduling with dependent setup times in textile industry. Comput. Ind. Eng. **174**, 108736 (2022). https://doi.org/10.1016/j.cie.2022.108736
6. Nie, Y., Biegler, L.T., Wassick, J.M.: Integrated scheduling and dynamic optimization of batch processes using state equipment networks. AIChE J. **58**(11), 3416–3432 (2012). https://doi.org/10.1002/aic.13738
7. Graham, R.L., Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G.: Optimization and approximation in deterministic sequencing and scheduling: a survey. In: Hammer, P.L., Johnson, E.L., Korte, B.H. (eds.) Discrete Optimization II. Annals of Discrete Mathematics, vol. 5, pp. 287–326. Elsevier (1979). https://doi.org/10.1016/S0167-5060(08)70356-X
8. Allahverdi, A., Ng, C.T., Cheng, T.C.E., Kovalyov, M.Y.: A survey of scheduling problems with setup times or costs. Eur. J. Oper. Res. **187**(3), 985–1032 (2008). https://doi.org/10.1016/j.ejor.2006.06.060
9. Kaban, A.K., Othman, Z., Rohmah, D.S.: Comparison of dispatching rules in job-shop scheduling problem using simulation: a case study. Int. J. Simul. Model. **11**(3), 129–140 (2012). https://doi.org/10.2507/IJSIMM11(3)2.201
10. Floudas, C.A., Lin, X.: Mixed integer linear programming in process scheduling: modeling, algorithms, and applications. Ann. Oper. Res. **139**(1), 131–162 (2005). https://doi.org/10.1007/s10479-005-3446-x
11. Wilson, J.M.: Alternative formulations of a flow-shop scheduling problem. J. Oper. Res. Soc. **40**(4), 395–399 (1989). https://doi.org/10.2307/2583256
12. Manne, A.S.: On the job-shop scheduling problem. Oper. Res. **8**(2), 219–223 (1960). https://doi.org/10.1287/opre.8.2.219
13. Pan, C.H.: A study of integer programming formulations for scheduling problems. Int. J. Syst. Sci. **28**(1), 33–41 (1997). https://doi.org/10.1080/00207729708929360
14. Frühwirth, T., Abdennadher, S.: Essentials of Constraint Programming. Springer, Berlin Heidelberg (2003). https://doi.org/10.1007/978-3-662-05138-2
15. Glover, F.: Future paths for integer programming and links to artificial intelligence. Comput. Oper. Res. **13**(5), 533–549 (1986). https://doi.org/10.1016/0305-0548(86)90048-1
16. Fonseca, G.H.G., Figueiroa, G.B., Toffolo, T.A.M.: A fix-and-optimize heuristic for the unrelated parallel machine scheduling problem. Comput. Oper. Res. **163**, 106504 (2024). https://doi.org/10.1016/j.cor.2023.106504
17. Ploskas, N.: Parameter tuning of linear programming solvers. In: 2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), pp. 1–6 (2022). https://doi.org/10.1109/SEEDA-CECNSM57760.2022.9933002
18. Rossi, A., Tiacci, L., Simonetti, M.: A mixed integer programming model to optimize production planning in the luxury textile industry. Procedia Comput. Sci. **253**, 1175–1184 (2025). https://doi.org/10.1016/j.procs.2025.01.179
19. Avgerinos, I., Mourtos, I., Vatikiotis, S., Zois, G.: Weighted tardiness minimization for unrelated machines with sequence-dependent and resource-constrained setups. arXiv preprint arXiv:2307.06671 (2023)
20. Williams, R., Pitts, S., Kamery, R.: Production scheduling problems for a totally integrated carpet manufacturer: a preliminary investigation (2004)



On the Use of Range Volatility Estimators for Rough Fractional Stochastic Volatility Model

Joana Sendas[1], Rui Pascoal[2], and Humberto Rocha[2(✉)]

[1] Faculty of Economics, University of Coimbra, Av Dias da Silva 165, 3004-512 Coimbra, Portugal

[2] Faculty of Economics, CeBER, University of Coimbra, Av Dias da Silva 165, 3004-512 Coimbra, Portugal

{ruiapsp,hrocha}@fe.uc.pt

Abstract. Financial models have faced challenges in accurately reproducing the dynamics of volatility. In particular, neither local volatility nor traditional stochastic volatility models are able to simultaneously produce realistic dynamics for the implied volatility surface and achieve a satisfactory fit to its shape. As a result, a new approach has emerged in mathematical finance, called rough volatility models, where the volatility dynamics of financial assets are modeled as a function of fractional Brownian motion with a Hurst exponent H smaller than one half. Rough volatility models are indeed capable of replicating the intended stylized facts using only a limited number of parameters. The Hurst index is estimated using realized volatility measurements as a proxy for the spot volatility. However, several authors argue that this estimator for volatility may be noisy due to the influence of market microstructure noise and that obtaining the necessary data for its calculation is very challenging, as it requires high-frequency data. Alternative proxies, novel in the context of rough volatility estimation, are proposed in this study (This study stems from a master's dissertation archived in the Scientific Repository of the University of Coimbra [1].) for the unobserved volatility process. The chosen proxies are range-based estimators: Open-to-Close, Rogers and Satchell, Parkinson, Garman and Klass, and Meilijson. These proxies will allow us to better determine whether volatility exhibits rough behavior and whether it can continue to reproduce the stylized facts, even in a low-frequency data framework. In addition, we will evaluate which of these estimators are best suited for this purpose.

Keywords: Rough Volatility · Fractional Brownian Motion · Range-based Estimators · Volatility Forecasting

1 Introduction

In financial mathematics, there is a constant effort to predict the uncertainty of future market behavior, despite it being a very challenging task, if not impossible.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 332–350, 2027.
https://doi.org/10.1007/978-3-032-30521-3_21

Volatility, in particular, is well-known for its significant repercussions, especially during financial crises, due to its central role in risk management, asset pricing, and portfolio allocation. Despite these challenges, significant achievements have been made in developing volatility models that capture market dynamics with increasing accuracy.

This study focus on a notable class of models called rough volatility models. This class of volatility models has gained prominence due to their superior ability to fit both historical and implied volatility, capable of reproducing stylized facts of financial markets. These models are characterized by their irregular paths, achieved by replacing the standard Brownian motion in volatility models with fractional Brownian motion (fBm), where the Hurst exponent is smaller than one half.

To estimate the Hurst parameter, rough volatility models require proxies for spot volatilities because volatility is not directly observed. For this purpose, most of these models, including the pioneering work of Gatheral *et al.* [2], which introduced the Rough Fractional Stochastic Volatility (RFSV) model, used realized volatility. However, realized volatility requires high- or ultra-high-frequency observations, which are often difficult to obtain. In a low-frequency data context, such approximations are not feasible. On the other hand, some authors claim that this proxy is subject to microstructure effects and turns out to be noisy. This suggests that the rough behavior of volatility could be influenced by the choice of realized volatility as a proxy for spot volatility.

There are several variations within this class of volatility models. In this paper we will focus on estimating the RFSV model using the Hurst parameter estimation method introduced in Gatheral *et al.* [2]. This model is recognized as the pioneering rough volatility model that laid the foundation for subsequent developments in the field, and is notable for its original contributions and wide acceptance in the literature. Furthermore, it is particularly well suited to effectively capture essential stylized facts. We aim to help answer the question: could rough volatility be an artifact of high-frequency return data? To address this question, we propose alternative proxies for the unobserved volatility, named range-based estimators. These estimators are shown to be robust to microstructure noise and efficient even with low-frequency data.

This paper is structured into four sections more, as outlined below. Section 2 covers fundamental concepts required to comprehend rough volatility models, introduces the RFSV model and presents the concepts of realized variance and range-based variance estimators. Section 3 focuses on empirical calculations, diagnostics, and interpretation of the results. In Sect. 4, we forecast volatility and evaluate the accuracy of each model. Lastly, in Sect. 5 the main conclusions of this study are summarized.

2 Methodology

Volatility estimation has been a central topic in financial mathematics literature for the past several decades. It is imperative to acknowledge the pivotal role of

Brownian motion when discussing volatility estimation. Indeed, the estimation of volatility has long been intricately associated with Brownian motion, particularly in Stochastic Volatility (SV) models, due to its ability to capture the stochastic nature of volatility. The Brownian motion is based on a phenomenon observed by Robert Brown in 1827, who noticed a ubiquitous, irregular motion of pollen grains suspended in water. Wiener rigorously formalized Brownian motion, establishing a framework for continuous stochastic processes, which is why Brownian motion is also referred as Wiener process [3].

Another strand of literature that has been heavily used in the field of financial mathematics is Stochastic Volatility models, which treat volatility as a time-varying process, more precisely, a continuous Brownian semi-martingale [2]. Gatheral *et al.* empirically demonstrated that log-volatility behaves essentially as a fractional Brownian motion with a Hurst exponent of approximately 0.1, indicating that volatility is rough [2]. They estimated the Hurst exponent using daily realized variance estimates as proxies for daily spot volatility. This led to the development of the Rough Fractional Stochastic Volatility (RFSV) model, which is characterized by a Hurst exponent $H < 1/2$.

Key concepts required to establish a comprehensive foundation for the RFSV model are detailed next.

2.1 Basic Concepts

Monofractal Scaling Property. The term *fractal* was introduced by Mandelbrot in 1975. A fractal can be defined as a "shape made of parts that are similar to the whole in some way" [4]. This section provides a concise overview of a class of self-affine processes called monofractals. Let $X(t)$ be a random process where $X(0) = 0$. Then $X(t)$ is considered self-affine, with parameter H if it satisfies the scaling rule

$$X(ct) \overset{d}{=} c^H X(t),$$

where $0 < H < 1$ is the Hurst exponent, and c and t are positive [5]. The Hurst exponent, also referred to as the self-affinity index, characterizes the long-term memory properties of time series.

The multiscaling properties of the self-affine function $X(t)$ can be investigated by calculating the structure function, which is defined as the q-th order moments of the increments $X(t + \Delta) - X(t)$:

$$S_q(\Delta) = \mathbb{E}\left[|X(t + \Delta) - X(t)|^q\right] \sim \Delta^{\zeta(q)},$$

where $q > 0$ and ζ_q is the scale invariant structure function exponent [6].

A self-affine process with a single Hurst exponent has $\zeta(q) = H_q$. These processes are called monofractal or uniscaling because $\zeta(q)$ scales as a linear function of order q [7]. Note that in these processes, H_q changes continuously with q, and H can be easily obtained from the slopes of log-log plots of S_p versus Δ for any q.

Fractional Brownian Motion. In 1923, Wiener defined ordinary Brownian motion W_t, also known as the Wiener process, where $t \in \mathcal{R}$ designates time, as a stochastic process that satisfies the following properties [3]: W_t is a continuous function of time with $W_0 = 0$; the increments $W_{t_2} - W_{t_1}$ are Gaussian with mean zero and variance $|t_2 - t_1|$, which implies that the increments are stationary; the increments over disjoint time intervals are independent for all $t_1 < t_2 < t_3 < t_4 < \ldots < t_m$.

Given the limitations of the standard Brownian motion model in capturing certain dynamics of financial markets, more sophisticated models have emerged. One of the most notable is fractional Brownian motion (fBm). Mandelbrot and Van Ness defined fBm, denoted by $\{W_t^H\}_{t\geq 0}$, in terms of Wiener integrals by

$$W_t^H = \int_0^t (t-s)^{H-1/2} dW_s + \int_{-\infty}^0 (t-s)^{H-1/2} - (-s)^{H-1/2} dW_s,$$

where $W_0^H = 0$ and $\{W_s\}_{s\in\mathcal{R}}$ denotes the standard Brownian motion, with $0 < H < 1$ as the Hurst parameter [8]. Note that if $H = 1/2$, the fractional Brownian motion W_t^H reduces to the standard Brownian motion.

A fBm, characterized as a centered Gaussian process $\{W_t^H\}_{t\geq 0}$ with parameter $0 < H < 1$, exhibits the following covariance function [9]:

$$\mathrm{E}\left[W_t^H W_s^H\right] = \frac{1}{2}\left(t^{2\,H} + s^{2\,H} - |t-s|^{2\,H}\right). \tag{1}$$

The covariance of increments, known as fractional Gaussian noise (fGn), is expressed based on Eq. (1) as follows:

$$\begin{aligned}\mathrm{E}\left[\left(W_{t_1}^H - W_{s_1}^H\right)\left(W_{t_2}^H - W_{s_2}^H\right)\right] = \\ = \frac{1}{2}\left(|t_1 - s_2|^{2\,H} + |t_2 - s_1|^{2\,H} - |t_2 - t_1|^{2\,H} - |s_2 - s_1|^{2\,H}\right). \end{aligned} \tag{2}$$

Based on Eq. (2), it is concluded that the increments are stationary, as they are Gaussian, centered, and share the same covariance structure.

Another significant property of fractional Brownian motion is self-affinity, more specifically, H-self-affinity. It can be observed that, for a fixed $a > 0$, the process $Z_t = W_{at}^H$ has the same covariance structure, and therefore the same distribution, as $a^H W_t^H$.

The fBm for $H \in (0, 1/2)$ is antipersistent, meaning that past increments are negatively correlated. On the other hand, the fBm for $H \in (1/2, 1)$ exhibits persistence, where past increments are positively correlated.

In the case where $H > 1/2$, the increments of the fBm exhibit long memory, in the sense that they display a slow decay in the autocorrelation function [2]:

$$\sum_{k=0}^{+\infty} \mathrm{Cov}\left[W_1^H, W_k^H - W_{k-1}^H\right] = +\infty.$$

Hence, the parameter H characterizes the long-memory properties of the data, as the autocorrelation function of the increments is of order k^{2H-2} as k approaches infinity.

As mentioned previously, the increments of fBm exhibit the properties of self-affinity and stationarity; therefore:

$$\mathbb{E}\left[\left|W_t^H - W_{t+\Delta}^H\right|^q\right] = K_q \Delta^{qH},$$

where K_q is the moment of order q of the absolute value of a standard Gaussian variable.

Realized Variance. Realized volatility gained significant recognition with the pioneering work of Andersen and Bollerslev [10], along with the growing availability of high-frequency data.

For simplicity, in addressing realized volatility, a non-parametric, model-free estimator, we assume a continuous-time diffusive setting, ignoring price jumps, and a frictionless market. Under these assumptions, the asset's logarithmic price process p is a semi-martingale to prevent arbitrage opportunities [11]. Its definition is given by

$$dp_t = \mu_t dt + \sigma_t dW_t, \quad 0 \leq t \leq T,$$

where μ_t is a predictable drift process, $\sigma_t > 0$ denotes a predictable spot volatility process and W_t is a standard Brownian motion. It is assumed that $E(\int_0^t \sigma_s^2 ds) < \infty$ [12]. The continuously compounded return throughout the period from $t-k$ and t, where $0 < k \leq t$, can be expressed as [12]:

$$r(t,k) = p(t) - p(t-k) = \int_{t-k}^{t} \mu_\tau \, d\tau + \int_{t-k}^{t} \sigma_\tau \, dW_\tau.$$

Thus, its quadratic variation (QV) is given by

$$QV(t,k) = \int_{t-k}^{t} \sigma^2(\tau) d\tau. \tag{3}$$

The expression for the realized variance (RV) of the logarithmic price process is given by

$$RV(t,k;n) = \sum_{j=1}^{n \cdot k} r\left(t-k+\frac{j}{n}, \frac{1}{n}\right)^2,$$

where $\{t-k+\frac{j}{n}, j = 1, ..., n \cdot k\}$ is a partition of the interval $[t-k, t]$ [12].

Due to the semimartingale properties, we have

$$RV(t,k;n) \to QV(t,k) \text{ as } n \to \infty.$$

The realized variance provides remarkable measures of actual market volatility [12]. However, although high-frequency data is becoming more accessible, it may not be available for some assets or only available for shorter time periods.

Range-Based Variance Estimators. Traditional estimators based solely on daily closing prices are often imprecise, which led to the development of alternative methods to improve accuracy. One notable approach is range-based volatility estimators, which estimate the volatility of a financial asset by considering the range of prices within a given time period, including the high, low, open, and close prices [13].

Let's assume that the price P follows a geometric Brownian motion and the log-price $p = \ln(P)$ follows a Brownian motion with zero drift and a diffusion parameter σ [14], i.e.,

$$dp_t = \sigma dB_t. \tag{4}$$

Let O_t represent the opening price, C_t the closing price, H_t the highest price, and L_t the lowest price of the day t [14]. The returns for open-to-close, open-to-high, and open-to-low are calculated by

$$c_t = \ln(C_t) - \ln(O_t),$$

$$h_t = \ln(H_t) - \ln(O_t),$$

$$l_t = \ln(L_t) - \ln(O_t).$$

The daily return c is a random variable drawn from a normal distribution with zero mean and variance σ^2. Since c^2 is an unbiased estimator of σ^2, the first range-based variance estimator [14], which in this work we will refer to as Open-to-Close variance for identification purposes, is given by

$$\widehat{\sigma_s^2} = c_t^2. \tag{5}$$

Since this estimator is highly noisy, Parkinson proposes the following estimator for variance [15]:

$$\widehat{\sigma_P^2} = \frac{(h_t - l_t)^2}{4 \ln 2}. \tag{6}$$

This estimator, now known as the Parkinson estimator, was the first to formalize the use of the range for volatility estimation [14].

Inspired by Parkinson's work, Garman and Klass [13] introduced an estimator that incorporates not only the high-low range but also the closing price, resulting in a more efficient estimator:

$$\widehat{\sigma_{GK}^2} = 0.511(h_t - l_t)^2 - 0.019\left(c_t(h_t + l_t) - 2h_t l_t\right) - 0.383c_t^2.$$

Garman and Klass also proposed a simplified minimum variance analytical estimator, given by:

$$\widehat{\sigma_{GK}^2} = 0.5(h_t - l_t)^2 - (2\ln 2 - 1)c_t^2. \tag{7}$$

Meilijson introduced a different estimator with a smaller variance than the Garman and Klass estimator [16]. This estimator is formulated as

$$\widehat{\sigma_M^2} = 0.274\sigma_1^2 + 0.16\sigma_s^2 + 0.365\sigma_3^2 + 0.2\sigma_4^2 \tag{8}$$

knowing that

$$\sigma_1^2 = 2\left[(h'_t - c'_t)^2 + (l'_t)^2\right],$$
$$\sigma_3^2 = 2\,(h'_t - c'_t - l'_t)\,c'_t,$$
$$\sigma_4^2 = -\frac{(h'_t - c'_t)\,l'_t}{2\ln 2 - 5/4},$$

where if $c_t > 0$, then $c'_t = c_t$, $h'_t = h_t$, and $l'_t = l_t$; and if $c_t < 0$, then $c'_t = -c_t$, $h'_t = -l_t$, and $l'_t = -h_t$. This condition implies that two different formulas are applied depending on whether c_t is positive or negative.

Rogers and Satchell proposed the following estimator that accommodates arbitrary drift [17]:

$$\widehat{\sigma^2_{RS}} = h_t(h_t - c_t) + l_t(l_t - c_t). \tag{9}$$

There are other estimators, but in this work, we address these due to their significant contributions to the literature and the notorious practical applications in improving volatility estimation accuracy.

2.2 Rough Fractional Stochastic Volatility

Background and Motivation. One of the most recognized stochastic volatility (SV) models is the Hull and White model which provides an option pricing framework where the volatility of the underlying asset is time-varying and correlated with a specific risk parameter [18]. Moreover, the long-memory property of the volatility process has been widely recognized as a stylized fact [19]. Comte and Renault expanded the Hull and White model to capture volatility persistence [20]. They proposed a continuous-time long-memory SV model, called Fractional Stochastic Volatility (FSV), by replacing the standard Brownian motion in the original model with a fBm with a Hurst exponent $H > 1/2$.

The concept of long memory in volatility, initially described as a slow decay of autocorrelation, slower than exponential decay, has become more precise, meaning that the autocorrelation function is not integrable or, more specifically, that it decays as a power law with an exponent less than 1 [2].

Another challenge that volatility models face is accurately replicating the volatility skew observed in the market [21]. Let $\sigma_{\mathrm{BS}}(k,\tau)$ denote the implied volatility, with log-moneyness k defined as $k = \ln\left(\frac{K}{S}\right)$, where K is the strike price, S is the current price of the underlying asset, and τ represents the time to maturity. The term structure of at-the-money (ATM) volatility skew is defined by [2]:

$$\psi(\tau) = \left|\frac{\partial}{\partial k}\sigma_{\mathrm{BS}}(k,\tau)\right|_{k=0}.$$

Fukasawa demonstrated that a SV model, in which volatility is driven by a fBm with Hurst exponent H, generates an ATM volatility skew of the form $\psi(\tau) \sim \tau^{H-1/2}$, at least for a small time to expiration τ [22]. To produce a volatility skew consistent with the market observations, the value of H must be

close to zero, as the skew term structure in financial markets behaves approximately as $\tau^{-1/2}$ for small τ. Gatheral *et al.* calculated non-parametric estimates of the S&P 500 index ATM volatility skews for a single day and verified that these estimates followed a power-law pattern fitting the form $\psi(\tau) = A\tau^{-0.4}$ [2].

Motivated by these observations, Gatheral *et al.* developed the Rough FSV (RFSV) model, considered a particular case of the FSV model that captures the roughness of the volatility process, rather than focusing on traditional long-memory behavior. The RFSV model, characterized by a Hurst exponent $H < 1/2$, reflects the irregular nature of the volatility process [2]. Their findings suggest that the dynamics of log-volatility are close to that of a fBm due to the Gaussianity and scaling properties of the increments of log-volatility, at any reasonable time scale. They estimated the Hurst parameter H to be approximately between 0.06 and 0.20, with a typical value around 0.1.

The Scaling Property and Normality. This section details the method proposed in [2] to estimate the Hurst exponent H of the RFSV model.

Assume access to discrete observations of the volatility process. For $q \geq 0$, the q-order moment of a log-volatility increment is defined as:

$$m(q, \Delta) = \frac{1}{N} \sum_{k=1}^{N} \left| \log\left(\sigma_{k\Delta}\right) - \log\left(\sigma_{(k-1)\Delta}\right) \right|^q,$$

where $\sigma_0, \sigma_\Delta, \ldots, \sigma_{k\Delta}, \ldots, \quad k \in \{0, N\}$, are discrete observations of the volatility process and $N = \lfloor T/\Delta \rfloor$ represents the total number of observations in the time interval $[0, T]$.

The fact that the Hurst parameter is considered a smoothness parameter arises from the following assumption that for some $s_q > 0$ and $b_q > 0$,

$$N^{qs_q} m(q, \Delta) \to b_q \tag{10}$$

as Δ tends to zero [2]. Rosenbaum demonstrated that s_q can be interpreted as the regularity of the volatility when measured in l_q norm [23]. If the logarithm of the volatility follows a fBm with Hurst parameter H, Eq. (10) hold in probability with $s_q = H$.

It is not possible to compute $m(q, \Delta)$ precisely, since the volatility process is not directly observable. Therefore, a proxy for spot volatility is required. Gatheral *et al.* applied realized volatility for this purpose. Assuming that the increments of the log-volatility process are stationary and that the law of large numbers applies, $m(q, \Delta)$ is an unbiased estimator of $\mathbb{E}\left[\left|\log\left(\sigma_\Delta\right) - \log\left(\sigma_0\right)\right|^q\right]$, and can consequently be viewed as its empirical counterpart. As a result, Gatheral *et al.* empirically discovered the following monofractal scaling relationship:

$$\mathbb{E}\left[\left|\log\left(\sigma_\Delta\right) - \log\left(\sigma_0\right)\right|^q\right] = b_q \Delta^{\zeta_q}, \tag{11}$$

where b_q denotes a constant and $\zeta_q = qs_q$ is the slope.

Additionally, Gatheral *et al.* made the important observation that the smoothness parameter s_q does not seem to depend on q. In fact, Gatheral *et al.*

revealed a linear relationship $\zeta_q \sim Hq$ by plotting ζ_q against q, with estimated H ranging between 0.06 and 0.20, leading to the conclusion that $H < 1/2$.

Andersen *et al.* reported in the early 2000s the well-established stylized fact that the distribution of the log-volatility process is very close to a Gaussian distribution [19]. Gatheral *et al.* confirmed that, for any Δ, the empirical distribution of log-volatility increments closely resembles a Gaussian distribution and that the 1-day fit of the normal density, scaled by Δ^H, aligns even more closely with the Gaussian distribution [2].

Model Specification. The increments of the log-volatility show a consistent scaling property with a constant smoothness parameter, and their distribution closely resembles a Gaussian distribution. Thus, a simple model is first proposed by [2]:

$$\log \sigma_{t+\Delta} - \log \sigma_t = v \left(W^H_{t+\Delta} - W^H_t \right), \tag{12}$$

where $t \geq 0$, W^H is a fBm with Hurst parameter H and v is a positive constant. The volatility can be expressed in exponential form:

$$\sigma_t = \sigma \exp\left(v W^H_t \right), \tag{13}$$

where σ is a positive constant.

This model does not exhibit the required stationarity, which can complicate both the mathematical analysis and the modeling process. To address this issue, the log-volatility is instead modeled as a fractional Ornstein-Uhlenbeck (fOU) process with a very long reversion timescale. The fOU process is an extension of the classical stationary Ornstein-Uhlenbeck process, and it also has a stationary solution [24].

The fOU process X_t is defined through the stochastic differential equation:

$$\mathrm{d}X_t = v \, \mathrm{d}W^H_t - \alpha \left(X_t - m \right) \mathrm{d}t, \tag{14}$$

and given by its solution:

$$X_t = v \int_{-\infty}^{t} e^{-\alpha(t-s)} \mathrm{d}W^H_t + m, \tag{15}$$

where $m \in \mathcal{R}$ is the long-term mean and $\alpha \in \mathcal{R}^+$ is the velocity of mean reversion.

Thus, the RFSV model is expressed as:

$$\sigma_t = \exp\left(X_t \right), \tag{16}$$

where $t \in [0, T]$ and X_t defined by Eq. (15) and $H < 1/2$.

Gatheral *et al.* verified that log-volatility modeled as a fOU behaves like a fBm as α tends to zero, through the following propositions [2]:

$$\mathbb{E}\left[\sup_{t \in [0,T]} \left| X^\alpha_t - X^\alpha_0 - v W^H_t \right| \right] \to 0, \tag{17}$$

and

$$\mathbb{E}\left[\left|X^{\alpha}_{t+\Delta} - X^{\alpha}_{t}\right|^{q}\right] \to v^{q} K_{q} \Delta^{qH}. \tag{18}$$

From Eq. (17), it is observed that if $\alpha \ll 1/T$, i.e., α is much less than $1/T$, the log-volatility process can be assumed to behave as a fBm within the interval of interest $[0, T]$. Additionally, from Eq. (18), the scaling property of a fBm is approximately reproduced by the fOU process for very small α.

As mentioned earlier, the RFSV model is considered a particular case of the FSV model. The Stochastic Differential Equation (SDE) of the RFSV model, as expressed in Eq. (16), is similar to that of the FSV model, with the fundamental difference being the range of the Hurst parameter H, where $H > 1/2$ in the FSV model and $H < 1/2$ in RFSV model.

Fukasawa demonstrated that a volatility process driven by a fBm leads to a term structure of ATM volatility skew of the form $\psi(\tau) \sim \tau^{H-1/2}$, at least for small τ [22]. This implies that in the FSV model, where $H > 1/2$, the ATM volatility skew increases with time to expiration τ for small values of τ. However, this is inconsistent with empirical observations, where the ATM volatility skew typically shows a decreasing pattern as τ increases. In contrast, the RFSV model, with $H < 1/2$, aligns with the observed term structure of ATM volatility skew [2,22]. Furthermore, in the RFSV model, $\alpha \ll 1/T$, indicating that the dynamics of the volatility process exhibit a behavior approximated to that of a fBm, as referenced in Eqs. (17) and (18).

Gatheral *et al.* challenge the stylized fact that volatility exhibits long memory. The classical concept of long memory is characterized by an autocorrelation function that decays as a power law. However, Gatheral *et al.* demonstrated that empirically, and within the RFSV model, the autocovariance function does not decay as a power law [2]. To further test this concept, Gatheral *et al.* applied some standard statistical procedures commonly used in financial econometrics to both the data and the RFSV model simulations to identify long memory [2]. Notably, the results confirm the fragility of long-memory estimation processes, as they falsely indicated that the RFSV model exhibits long memory, highlighting that traditional methods may mistakenly identify long memory. A possible justification for these spurious results may be the strict modeling assumptions of these procedures (See Sect. 4 of [2]).

The innovative aspect of our approach lies in the proxies used for spot volatility. Gatheral *et al.* calculate the Hurst exponent H using realized volatility. In our study, however, we will calculate the Hurst exponent using range-based volatility estimators.

3 Empirical Study

3.1 Data and Preliminary Analysis

In order to estimate the parameters of the RFSV model, it is necessary to use a proxy for spot volatility values, as it is not possible to observe the volatility process directly. Gatheral *et al.* used daily realized volatility estimates [2].

We propose an alternative approach to estimating the parameters of the RFSV model by using range-based volatility estimators as proxies for spot volatility values, and evaluate and compare their performance. The range-based estimators include Open-to-Close volatility, Garman and Klass volatility, Parkinson volatility, Rogers and Satchell volatility and Meilijson volatility.

The dataset consists of nine stock indices, including the Standard & Poor's 500 (S&P500), Financial Times Stock Exchange 100 (FTSE100), Nikkei 225 (N225), Deutscher Aktienindex (DAX), All Ordinaries (AO), Cotation Assistée en Continu (CAC40), Swiss Market Index (SMI), Índice Bursátil Español 35 (IBEX35), and Nifty 50 (N50). The period covered in this analysis spans from April 19, 2001, to September 26, 2016.[1]

The daily realized variance estimates are available from the Oxford-Man Institute of Quantitative Finance Realized Library, with realized variance computed using 5-minute intraday intervals. The range-based estimators are based on daily price data, specifically the open, close, high, and low prices that were collected for that time period. The number of observations[2] is slightly higher for the range-based estimators than for the realized variance, with a notable difference in the Nifty 50 index, as shown in Table 1.

Table 1. Number of observations for each index.

Estimators/Index	SP500	CAC40	SMI	IBEX35	N50	FTSE100	N225	DAX	AO
Range-based estimators	3884	3951	3894	3923	3842	3898	3784	3933	3925
Realized volatility	3863	3935	3864	3904	3509	3877	3736	3908	3856

The statistics of each estimator are evaluated to better understand its behavior. These include the mean, median, maximum, minimum, standard deviation, skewness, excess kurtosis, and finally, autocorrelation (with one lag). The statistics for the Meilijson estimator and the realized volatility are presented in Tables 2 and 3, respectively. From these, we can draw interesting insights. The median values are lower than the means, suggesting a right-skewed distribution, also known as a positively skewed distribution of the data. Excess kurtosis values are notably high, indicating the presence of heavy tails in the volatility distribution. The autocorrelation (with one lag) is consistently positive across all estimators and indices, which indicates volatility clustering, another well-known stylized fact in financial markets [25]. Similar results are obtained for the remaining estimators. The normality test Jarque-Bera was performed, and for all estimators, the null hypothesis of normality was rejected with a p-value

[1] During this period, the indices included the following numbers of companies: S&P500 had 500 companies, FTSE100 had 100, Nikkei225 had 225, DAX had 30, All Ordinaries had 500, CAC40 included 40, SMI had 20, IBEX35 had 35, and Nifty 50 had 50 companies.

[2] The term "observations" refers to the number of daily variance values. We are not referring to the high-frequency data used to compute realized variance, which is based on intraday information.

Table 2. Descriptive statistics of the daily Meilijson volatility.

Statistic/Index	SP500	CAC40	SMI	IBEX35	N50	FTSE100	N225	DAX	AO
Mean	0.0077	0.0098	0.0079	0.0106	0.0107	0.0084	0.0082	0.0108	0.0055
Median	0.0063	0.0082	0.0064	0.0092	0.0088	0.0069	0.0072	0.0088	0.0047
Maximum	0.0854	0.0670	0.0877	0.0534	0.1256	0.0586	0.1066	0.0846	0.0466
Minimum	0.0012	0.0012	0.0014	0.0016	0.0017	0.0014	0.0010	0.0011	0.0008
Standard Dev.	0.0057	0.0063	0.0055	0.0062	0.0076	0.0057	0.0052	0.0074	0.0034
Skewness	3.6567	2.4387	3.3786	1.8840	4.5152	2.8019	4.3101	2.5514	2.8668
Excess Kurtosis	21.7898	7.0062	19.2253	3.0057	37.7419	10.2025	44.5314	8.0326	13.7861
Autocorrelation	0.6906	0.6841	0.6857	0.6817	0.6055	0.6047	0.5355	0.7174	0.6047

Table 3. Descriptive statistics of the daily realized volatility.

Statistic/Index	SP500	CAC40	SMI	IBEX35	N50	FTSE100	N225	DAX	AO
Mean	0.0086	0.0105	0.0081	0.0109	0.0105	0.0077	0.0093	0.0114	0.0061
Median	0.0071	0.0090	0.0066	0.0100	0.0088	0.0063	0.0082	0.0095	0.0052
Maximum	0.0965	0.0675	0.0529	0.0690	0.1005	0.0571	0.0619	0.0802	0.0394
Minimum	0.0015	0.0020	0.0026	0.0020	0.0016	0.0020	0.0025	0.0021	0.0011
Standard Dev.	0.0060	0.0061	0.0048	0.0059	0.0066	0.0050	0.0051	0.0070	0.0036
Skewness	3.5936	2.5244	2.7264	2.1519	3.6433	2.7175	3.1835	2.7694	2.6560
Excess Kurtosis	21.2837	8.4299	8.4967	7.2521	22.5662	9.3577	15.3124	10.3073	8.8260
Autocorrelation	0.8306	0.8183	0.8588	0.7970	0.7946	0.8405	0.7431	0.8422	0.7059

smaller than 2×10^{-16}, as expected due to the presence of skewness, heavy tails, and excess kurtosis.

3.2 Parameters Estimations

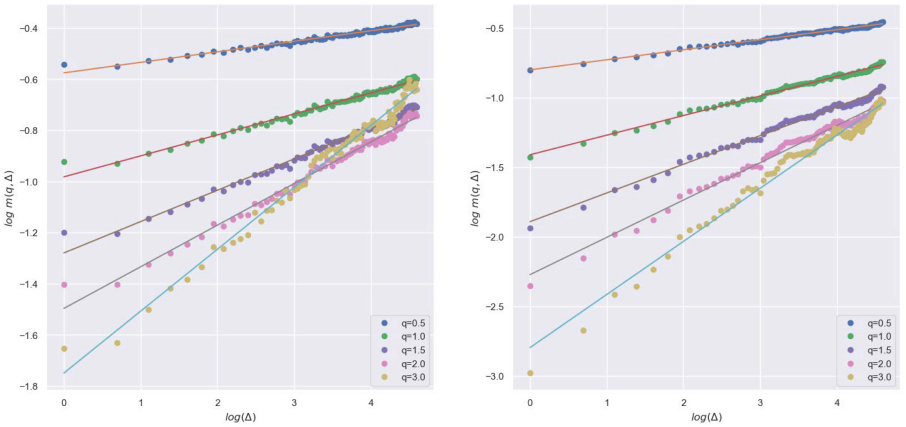
The parameters of the RFSV model are estimated by showing that log-volatility exhibits the scaling property (12) and follows a distribution similar to the normal distribution, using range-based estimators as proxies.

The computation of $m(q, \Delta)$ based on non-observable volatility is not possible in practice, so we use as proxies the volatility estimators calculated. We compute $m(q, \Delta)$ for different values of q. For illustration purposes, considering the Meilijson estimator and the realized variance, in Fig. 1, we observe that the values of $\log m(q, \Delta)$, plotted against $\log \Delta$ for each q, are closely clustered around their respective linear regression lines. The regression lines with lower slopes exhibit a better fit, where ζ_q represents the slope. Additionally, we note that the plots show quite different slopes for each estimator, however, all seem to exhibit the monofractal scaling property described in Eq. (11), leading to the estimation of H, as will be shown in the next step. Table 4 shows the slope ζ_q values. The Open-to-Close estimator shows the lowest values, which corresponds to lines almost horizontal in a plot $\log m(q, \Delta)$ against $\log \Delta$. The realized estimator shows higher values, indicating that it is more sensitive to changes in the time scale.

We can estimate parameter ν. Noting that $m(2, \Delta)$ can be seen as the empirical counterpart of $E\left[\left|\log(\sigma_{t+\Delta}) - \log(\sigma_t)\right|^2\right]$, the parameter ν is obtained from

Table 4. Values of ζ_q for each estimator for the S&P 500 index.

Estimators/q	0.5	1	1.5	2	3
Open-to-Close	0.010	0.0188	0.0261	0.0325	0.0432
Garman and Klass	0.0397	0.0798	0.1199	0.1600	0.2397
Parkinson	0.0406	0.0811	0.1213	0.1612	0.2400
Rogers and Satchell	0.0254	0.0481	0.0680	0.0851	0.1133
Meilijson	0.0412	0.0820	0.1225	0.1626	0.2419
Realized estimator	0.0720	0.1410	0.2063	0.2682	0.3818



(a) Meilijson estimator. (b) Realized estimator.

Fig. 1. $\log m(q, \Delta)$ as a function of $\log \Delta$, S&P 500.

Eq. (18), where, setting $q = 2$, we deduce that

$$m(2, \Delta) \approx \nu^2 \Delta^{2H}.$$

By analyzing Table 5, we observe that the Hurst exponent values vary across the proxies used, yet they consistently remain below 1/2. The parameter ν, often referred to as the "volatility of volatility," appears notably higher when employing the Open-to-Close estimator, indicating larger fluctuations in the variance process. This, combined with the lower value of H, suggests a model exhibiting more irregular behavior. Conversely, when realized volatility is used as the proxy, the model shows a higher H and a smaller ν, implying smoother swings in the variance process and a less irregular volatility trajectory.

Additionally, we observe that the SMI index shows the highest H in our sample, while the All Ordinaries index presents the lowest H, except for the Open-to-Close and Garman and Klass estimators. Under the Open-to-Close estimator, the All Ordinaries index still holds the lowest H value, while the S&P 500 index displays the highest. For the Garman and Klass estimator, the SMI retains the highest H.

Table 5. Estimated H for all indices.

H	Index	ν
0.0164	S&P 500	1.4961
0.0104	CAC 40	1.6064
0.0148	SMI	1.5173
0.0102	IBEX 35	1.5782
0.0106	Nifty 50	1.5534
0.0146	FTSE 100	1.5281
0.0105	Nikkei 225	1.5986
0.0128	DAX	1.5766
0.0089	All Ordinaries	1.5230

Open-to-Close estimator

H	Index	ν
0.0791	S&P 500	0.4792
0.0786	CAC 40	0.4529
0.0937	SMI	0.4171
0.0750	IBEX 35	0.4329
0.0697	Nifty 50	0.4725
0.0795	FTSE 100	0.4505
0.0630	Nikkei 225	0.4957
0.0880	DAX	0.4442
0.0542	All Ordinaries	0.4814

Garman and Klass estimator

H	Index	ν
0.0418	S&P 500	0.6770
0.0600	CAC 40	0.5455
0.0727	SMI	0.5135
0.0547	IBEX 35	0.5323
0.0523	Nifty 50	0.5383
0.0463	FTSE 100	0.6381
0.0339	Nikkei 225	0.6570
0.0662	DAX	0.5474
0.0142	All Ordinaries	0.8731

Rogers and Satchell estimator

H	Index	ν
0.0799	S&P 500	0.5029
0.0741	CAC 40	0.4852
0.0868	SMI	0.4537
0.0688	IBEX 35	0.4700
0.0658	Nifty 50	0.5075
0.0851	FTSE 100	0.4593
0.0619	Nikkei 225	0.5301
0.0850	DAX	0.4734
0.0546	All Ordinaries	0.5158

Parkinson estimator

H	Index	ν
0.0813	S&P 500	0.4732
0.0800	CAC 40	0.4492
0.0946	SMI	0.4140
0.0762	IBEX 35	0.4294
0.0689	Nifty 50	0.4735
0.0820	FTSE 100	0.4435
0.0654	Nikkei 225	0.4865
0.0889	DAX	0.4394
0.0575	All Ordinaries	0.4749

Meilijson estimator

H	Index	ν
0.1335	S&P 500	0.3236
0.1325	CAC 40	0.2933
0.1817	SMI	0.2219
0.1271	IBEX 35	0.2868
0.1181	Nifty 50	0.3119
0.1449	FTSE 100	0.2688
0.1165	Nikkei 225	0.3214
0.1518	DAX	0.2751
0.0880	All Ordinaries	0.3504

Realized estimator

4 Forecasting Volatility Using RFSV

We would like to find out which of the estimators is predicted most accurately by the RFSV model. For this purpose, we compare the variance and volatility forecasts of several indexes obtained using the RFSV model, with the parameters estimated for each estimator.

The log-volatility process can be approximated as a fBm, over any reasonable timescale of interest, as shown by Eq. (17) and Eq. (18). Therefore, we obtain an approximate non-stationary model for the increments of the log-volatility in Eq. (12), which can be rewritten as:

$$\log \sigma_t^2 \approx 2\nu W_t^H + C,$$

for some constant C [2].

The prediction method is based on the following formula:

$$\mathbb{E}\left[W_{t+\Delta}^H \mid \mathcal{F}_t\right] = \frac{\cos(H\pi)}{\pi}\Delta^{H+1/2}\int_{-\infty}^{t}\frac{W_s^H}{(t-s+\Delta)(t-s)^{H+1/2}}ds,$$

where W^H denotes a fBm with $H < 1/2$ and $\mathcal{F}_t$ the filtration it generates [2].

By substituting the fBm with the log-volatility, we have the prediction formula for the log-variance:

$$\mathbb{E}\left[\log\sigma_{t+\Delta}^2 \mid \mathcal{F}_t\right] = \frac{\cos(H\pi)}{\pi}\Delta^{H+1/2}\int_{-\infty}^{t}\frac{\log\sigma_s^2}{(t-s+\Delta)(t-s)^{H+1/2}}ds, \quad (19)$$

where $\Delta = 1, 5, 20$ [2]. Note that we want the formula for the RFSV predictor of the variance. Nuzman and Poor showed that $W_{t+\Delta}^H$is conditionally Gaussian with the following conditional variance [26]:

$$\mathrm{Var}[W_{t+\Delta}^H|\mathcal{F}_t] = c\Delta^{2H}$$

where

$$c = \frac{\Gamma(3/2-H)}{\Gamma(H+1/2)\Gamma(2-2H)}.$$

Thus, we easily arrive at the formula for the RFSV predictor of the variance [2]:

$$\mathbb{E}\left[\sigma_{t+\Delta}^2 \mid \mathcal{F}_t\right] = \exp\left\{\mathbb{E}\left[\log\sigma_{t+\Delta}^2 \mid \mathcal{F}_t\right] + 2c\nu^2 2H\right\},$$

where $\mathbb{E}\left[\log\sigma_{t+\Delta}^2 \mid \mathcal{F}_t\right]$ is defined by Eq. (19) and approximated through a Riemann sum.

We need a reliable statistic to compare the forecast quality of the estimators. For this purpose, we compute the following ratio P, which compares the mean squared error of our predictor with the approximate variance of the log-variance:

$$P = \frac{\sum_{k=500}^{N-\Delta}\left(\log(\sigma_{k+\Delta}^2) - \widehat{\log(\sigma_{k+\Delta}^2)}\right)^2}{\sum_{k=500}^{N-\Delta}\left(\log(\sigma_{k+\Delta}^2) - \mathcal{E}[\log(\sigma_{t+\Delta}^2)]\right)^2},$$

where $\mathcal{E}[\log(\sigma_{t+\Delta}^2)]$ represents the empirical mean of the log-variance over the entire time period and $\widehat{\log(\sigma_{k+\Delta}^2)}$ is the predictor of $\log(\sigma_{k+\Delta}^2)$ [2].

The ratio P measures the proportion of the total variance of $\log(\sigma_{k+\Delta}^2)$ that is not captured by the predicted values $\widehat{\log(\sigma_{k+\Delta}^2)}$. Ideally, a P value close to zero indicates better predictive performance, as it implies that most of the variance is captured by the model. Conversely, if $P = 1$, the mean squared error of the predictor matches the empirical variance of the log-volatility, indicating that the predictor does not improve upon a simple average. If P is greater than 1, it means that the model performs worse than simply using the expected mean.

Table 6 display the ratio P. It is clear that realized volatility, as a proxy, is more accurately captured in its behavior by the RFSV model compared to

Table 6. Ratio P for different volatility estimators proxies.

Index/Estimator	Open-to-close	Garman and Klass	Parkinson	Rogers and Satchell	Meilijson	Realized
S&P 500 $\Delta = 1$	1.9556	0.5643	0.5895	0.8863	**0.5439**	**0.3009**
S&P 500 $\Delta = 5$	1.9644	0.6661	0.6828	0.9457	**0.6628**	**0.4732**
S&P 500 $\Delta = 20$	2.1029	**0.8539**	0.8904	1.0644	0.8560	**0.7082**
CAC 40 $\Delta = 1$	2.2324	0.5450	0.6017	0.7102	**0.5427**	**0.3044**
CAC 40 $\Delta = 5$	2.2206	**0.6417**	0.6905	0.7793	0.6459	**0.4454**
CAC 40 $\Delta = 20$	2.3097	**0.8171**	0.8674	0.9304	0.8176	**0.6622**
SMI $\Delta = 1$	2.0751	**0.5332**	0.5965	0.7063	0.5338	**0.2221**
SMI $\Delta = 5$	2.0965	**0.6340**	0.6903	0.7744	0.6384	**0.3628**
SMI $\Delta = 20$	2.2118	**0.8350**	0.8952	0.9408	0.8382	**0.6046**
IBEX 35 $\Delta = 1$	2.1293	0.4643	0.5236	0.6244	**0.4604**	**0.2402**
IBEX 35 $\Delta = 5$	2.1158	0.5585	0.6124	0.6824	**0.5567**	**0.3669**
IBEX 35 $\Delta = 20$	2.1828	**0.6889**	0.7415	0.7866	0.6890	**0.5297**
Nifty 50 $\Delta = 1$	2.0751	0.5348	0.5730	0.6573	**0.5335**	**0.3030**
Nifty 50 $\Delta = 5$	2.0457	**0.6433**	0.6691	0.7400	0.6443	**0.4644**
Nifty 50 $\Delta = 20$	2.1232	0.8071	0.8340	0.8817	**0.8059**	**0.6514**
FTSE 100 $\Delta = 1$	2.0342	0.5369	0.5530	0.8165	**0.5265**	**0.2359**
FTSE 100 $\Delta = 5$	2.0463	0.6195	0.6355	0.8556	**0.6160**	**0.3614**
FTSE 100 $\Delta = 20$	2.1588	0.7881	0.8287	0.9853	**0.7846**	**0.5664**
Nikkei $\Delta = 1$	2.1918	0.6995	0.7222	0.9982	**0.6781**	**0.3975**
Nikkei $\Delta = 5$	2.1860	0.8294	0.8489	1.0489	**0.8140**	**0.5941**
Nikkei $\Delta = 20$	2.2485	**0.9941**	1.0186	1.1547	0.9953	**0.8417**
DAX $\Delta = 1$	2.1563	0.5449	0.6066	0.7057	**0.5428**	**0.2983**
DAX $\Delta = 5$	2.1357	**0.6443**	0.6805	0.7990	0.6484	**0.4543**
DAX $\Delta = 20$	2.2702	**0.8445**	0.8938	0.9727	0.8452	**0.7012**
All Ordinaries $\Delta = 1$	2.0559	0.5926	0.6319	1.1734	**0.5807**	**0.3850**
All Ordinaries $\Delta = 5$	2.0217	0.6530	0.6753	1.1629	**0.6441**	**0.4519**
All Ordinaries $\Delta = 20$	2.0694	0.7517	0.7907	1.1587	**0.7514**	**0.5959**

Note: Best and second-best predictions are in boldface.

range-based estimators. Among the range-based volatility estimators, the Meilijson estimator performs better, followed by the Garman-Klass. The high values observed in the Open-to-Close estimator show its significant level of noise. We also note that the predictive accuracy deteriorates with the forecast horizon, as expected, and that as Δ increases, the gap between the realized variance and the other estimators decreases.

Regarding the indices that are better predicted by these models, those depend on the value of Δ. For $\Delta = 1$, the best-performing index was the SMI for most estimators except for the Open-to-Close, where it was the S&P 500. Con-

versely, the worst-performing index varies more according to the estimator used as a proxy. For the Garman and Klass and Parkinson estimators, the worst-performing index was the IBEX 35, whereas for the Realized Variance and Rogers and Satchell estimators, it was the All Ordinaries. Additionally, for the Meilijson estimator, the worst was the Nikkei, and for the Open-to-Close, it was the CAC 40.

In general, as Δ increases, the indices that yield the best results remain consistent across all estimators, as do those yielding the worst results, except for the Open-to-Close estimator. For $\Delta = 5$, the best index for most estimators was the IBEX 35, except for the Open-to-Close, where it was the Nifty 50. The worst-performing index was the Nikkei, except for the Open-to-Close, where it was the S&P 500. Finally, for $\Delta = 20$, the best-performing index was the IBEX 35, and the worst was the Nikkei, except for the Open-to-Close estimator, where the best index was the S&P 500 and the worst was the CAC 40.

5 Conclusion

The RFSV model, known for its consistency with some financial stylized facts, has relatively few parameters. Nevertheless, the estimation of these parameters typically relies on the use of high-frequency realized variance, which can be costly and unavailable for some assets. To mitigate this, range-based volatility estimators are proposed that use daily open, high, low and close prices. Range-based proxies were found to be efficient in estimating the RFSV model parameters. We verified that the logarithm of range-based volatilities behaves like a fractional Brownian motion with a Hurst exponent less than 1/2. Furthermore, we observe that all Hurst exponents estimated using range-based estimators are smaller than those obtained from realized variance. This suggests that the RFSV model estimated with range proxies seems to produce values of H consistent with those obtained from the ATM volatility skew.

This study supports the idea that volatility is indeed rough, demonstrating this behavior even in low-frequency data. Furthermore, the RFSV model can be still useful, when calculated using range-based proxies. We believe that the Open-to-Close estimator should indeed not be used for this purpose. However, although realized variance shows better results, the Meilijson, Garman and Klass, and Parkinson estimators, in that order, perform relatively well and could serve as alternatives if high-frequency data are not accessible.

It is crucial for future research to investigate the performance of these range-based volatility estimators as proxies for the RFSV model when applied to other methods of estimating the Hurst exponent. This includes examining smoother versions of these estimators that incorporate more recent past observations alongside current ones, to better understand their limitations and potential improvements. We note a clear relationship between the indices and the quality of the results, which may be influenced by factors such as market capitalization, transaction volume, or market efficiency. We suggest that financial analysts further investigate these aspects to draw more conclusive insights.

Acknowledgments. This work has been funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., through project PTDC/MAT-APL/1286/2021 and project UID/05037/2025 with DOI https://doi.org/10.54499/UID/05037/2025.

References

1. Sendas, J.: Estimation of rough fractional stochastic volatility using range-based volatility estimators. Master's dissertation, University of Coimbra Scientific Repository (2024). https://estudogeral.uc.pt/handle/10316/118520
2. Gatheral, J., Jaisson, T., Rosenbaum, M.: Volatility is rough. Quant. Finance **18**, 933–949 (2018)
3. Wiener, N.: Differential-space. J. Math. Phys. **2**, 131–174 (1923)
4. Mandelbrot, B.B.: The Fractal Geometry of Nature. W.H. Freeman, New York, NY (1982)
5. Mandelbrot, B.B., Fisher, A.J., Calvet, L.E.: A multifractal model of asset returns. In: Cowles Foundation Discussion Paper, Cowles Foundation for Research in Economics. Yale University (1997)
6. Schmitt, F., Schertzer, D., Lovejoy, S.: Multifractal fluctuations in finance. Int. J. Theor. Appl. Finan. **3**, 361–364 (2000)
7. Chapman, S.C., Hnat, B., Rowlands, G., Watkins, N.W.: Scaling collapse and structure functions: identifying self-affinity in finite length time series. Phys. Rev. E **71**, 036114 (2005)
8. Mandelbrot, B.B., Van Ness, J.W.: Fractional Brownian motions, fractional noises and applications. SIAM Rev. **10**, 422–437 (1968)
9. Shevchenko, G.: Fractional Brownian motion in a nutshell. arXiv:1406.1956 (2014)
10. Andersen, T.G., Bollerslev, T.: Intraday periodicity and volatility persistence in financial markets. J. Empirical Finan. **4**, 115–158 (1997)
11. Back, K.: Asset pricing for general processes. J. Math. Econ. **20**, 371–395 (1991)
12. Andersen, T.G., Benzoni, L.: Realized Volatility. Working Paper, Federal Reserve Bank of Chicago (2008)
13. Garman, M.B., Klass, M.J.: On the estimation of security price volatilities from historical data. J. Bus. **53**, 67–78 (1980)
14. Molnár, P.: Properties of range-based volatility estimators. Finan. Res. Lett. **8**, 79–89 (2011)
15. Parkinson, M.: The extreme value method for estimating the variance of the rate of return. J. Bus. **53**, 61–65 (1980)
16. Meilijson, I.: The Gini index of a mixture of income distributions. Metron **67**, 85–97 (2009)
17. Rogers, L.C.G., Satchell, S.E.: Estimating variance from high, low and closing prices. Ann. Appl. Probab. **1**, 504–512 (1991)
18. Hull, J., White, A.: One-factor interest-rate models and the valuation of interest-rate derivative securities. J. Financ. Quant. Anal. **28**, 235–254 (1993)
19. Andersen, T.G., Bollerslev, T., Diebold, F.X., Ebens, H.: The distribution of realized stock return volatility. J. Financ. Econ. **61**, 43–76 (2001)
20. Comte, F., Renault, E.: Long memory in continuous-time stochastic volatility models. Math. Finan. **8**, 291–323 (1998)
21. Rebonato, R.: Volatility and Correlation: The Perfect Hedger and the Fox. Wiley, Chichester, UK (2004)

22. Fukasawa, M.: Asymptotic analysis for stochastic volatility: martingale expansion. Finance Stochast. **15**, 635–654 (2011)
23. Rosenbaum, M.: Estimation of the volatility persistence in a discretely observed diffusion model. Stoch. Process. Appl. **118**, 1434–1462 (2008)
24. Cheridito, P., Kawaguchi, H., Maejima, M.: Fractional Ornstein-Uhlenbeck processes. Electron. J. Probab. **8**, 1–14 (2003)
25. Cont, R.: Empirical properties of asset returns: stylized facts and statistical issues. Quant. Finance **1**, 223–236 (2001)
26. Nuzman, C.J., Poor, H.V.: Linear estimation of self-similar processes via Lamperti's transformation. J. Appl. Probab. **37**, 429–452 (2000)

A Reinforcement Learning–Guided General Variable Neighborhood Search for the Traveling Salesman Problem

Karina Assini Andreatta(✉), Maria Claudia Silva Boeres, and Renato Elias Nunes de Moraes

Federal University of Espírito Santo (UFES), Avenue Fernando Ferrari 514, Vitória, ES 29075-910, Brazil
karinassini@gmail.com, {maria.boeres,renato.moraes}@ufes.br
https://www.ufes.br

Abstract. This paper proposes RL-GVNS, a reinforcement learning–guided General Variable Neighborhood Search framework for the Traveling Salesman Problem (TSP), which introduces adaptivity in both the construction and intensification phases. The proposed RL-GVNS extends the classical GVNS by integrating a multi-agent reinforcement learning (MARL) initializer and an RL-driven local search, where neighborhood operators are selected online using Q-learning. Unlike standard GVNS, which relies on a fixed ordering of local search operators, RL-GVNS dynamically adapts its search behavior based on performance feedback, enabling an effective response to non-stationary search dynamics. Although all evaluated configurations, combining classical or MARL-based initialization with standard GVNS or RL-GVNS, perform similarly in small instances, learning-enhanced variants show superior robustness as instance size increases. In particular, the combined MARL + RL-GVNS configuration achieves lower average gaps with respect to the best-known solution, higher best-known solution hit rates, and, in most cases, comparable or reduced execution times relative to non-adaptive GVNS baselines.

Keywords: General Variable Neighborhood Search · Reinforcement Learning · Traveling Salesman Problem

1 Introduction

The Traveling Salesman Problem (TSP) is a classical benchmark in combinatorial optimization. Given the pairwise travel costs between cities, the goal is to find a minimum–cost Hamiltonian tour. Although simply defined, TSP is NP–hard: the search space grows factorially with the number of cities, making exact methods computationally impractical for medium and large–scale instances [23]. This limitation motivates the use of metaheuristic approaches, which yield high-quality solutions at practical computational cost [27]. Among these, Variable

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 351–368, 2027.
https://doi.org/10.1007/978-3-032-30521-3_22

Neighborhood Search (VNS), proposed in [26], leverages this trade-off by relying on a systematic exploration of multiple neighborhood structures. The underlying principle of VNS is that a solution that is locally optimal with respect to one neighborhood structure may cease to be optimal when evaluated under a different neighborhood. The method alternates between diversification and intensification by changing the neighborhood during the search process. Diversification is performed through a shaking phase, which applies controlled perturbations to move the search away from the current solution, while intensification is achieved by applying local search procedures within a selected neighborhood to refine the solution and reach a local optimum [21].

In General Variable Neighborhood Search (GVNS), intensification is commonly implemented through Variable Neighborhood Descent (VND) [13]. However, the fixed ordering of local search operators in standard VND leads to static search behavior, motivating adaptive strategies. Although GVNS has been successfully applied to computationally hard optimization problems that arise in practical settings, such as routing, scheduling, and network design [7,18,20,25,34], recent studies have increasingly explored the integration of machine learning techniques to guide metaheuristic decisions, including promising results obtained through reinforcement learning–based operator adaptation [19,29,33].

Despite recent progress, most approaches focus on population-based or evolutionary frameworks [2,17,30], or rely on end-to-end neural construction policies via deep reinforcement learning [22], which require substantial training data and do not readily generalize across instance distributions. Trajectory-based single-solution approaches to the TSP have largely retained fixed operator schemes [7,16,34], with little provision for adapting the selection of local search operators during the run. The underlying difficulty is that TSP instances exhibit non-stationary search landscapes: perturbation moves that reliably yield improvements in early iterations tend to lose their effectiveness as the algorithm approaches high-quality local optima, making a static ordering progressively less efficient as the search converges [21].

As shown in Table 1, the majority of the reviewed approaches incorporate some form of adaptability, although mainly at the population or initialization level. The method proposed here pursues a different direction, placing the adaptation mechanism within the local search phase itself. Q-learning is used to guide the sequencing of neighborhood operators throughout intensification, implementing an Adaptive Operator Selection (AOS) scheme [10] that adjusts operator priorities based on observed performance during the run. The shaking procedure was also modified: at each neighborhood level, a set of operators of increasing structural strength is defined, from which a move is sampled at random, replacing the standard single fixed perturbation and providing a more balanced coverage of the search space across shaking levels.

The remainder of the paper is structured as follows. Section 2 introduces the TSP and reviews the GVNS framework. Section 3 describes the proposed RL-GVNS method, including adaptive initialization, shaking strategy, and local

Table 1. Comparative analysis of related works. Columns indicate: (1) use of VNS or GVNS; (2) population-based approach; (3) explicit parameter calibration or control; (4) adaptive operator selection (AOS); (5) adaptive shaking mechanism; (6) availability of open-source implementation.

Author	(1)	(2)	(3)	(4)	(5)	(6)
da Silva et al. (2010) [7]	✓	✗	–	✗	✗	–
Queiroz dos Santos et al. (2014) [29]	✓	✗	–	✓	✗	–
Todosijevic et al. (2016) [33]	✓	✗	–	✓	✗	–
Menendez et al. (2017) [25]	✓	✗	–	✗	✗	–
Venkatesh et al. (2018) [34]	✓	✗	–	✗	✗	–
Hore et al. (2018) [16]	✓	✗	–	✗	✗	–
Alipour et al. (2018) [2]	✗	✓	–	✗	✗	–
Kool et al. (2019) [22]	✗	✗	–	✗	✗	✓
Karakostas et al. (2019) [20]	✓	✗	Partial	✗	✗	–
Karakostas et al. (2020) [18]	✓	✗	–	✗	✗	–
Karakostas et al. (2022) [19]	✓	✗	–	✓	✓	–
Ruan et al. (2024) [30]	✗	✓	–	✗	✗	–
Jeong et al. (2025) [17]	✗	✓	–	✓	✗	–
This work	✓	✗	✓	✓	✓	✓

search driven by reinforcement learning. Computational experiments and results are presented in Sect. 4. Finally, Sect. 5 summarizes the main findings and discusses directions for future work.

2 Preliminaries

2.1 Traveling Salesman Problem (TSP)

The Traveling Salesman Problem (TSP) is a canonical problem in combinatorial optimization and a standard benchmark for heuristic and metaheuristic algorithms [27]. In this work, the symmetric variant is considered and defined on a complete weighted graph $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ represents the set of cities and each edge $(i, j) \in E$ is associated with a non-negative travel cost c_{ij}, with $c_{ij} = c_{ji}$ for all $i, j \in V$.

A feasible solution is a Hamiltonian cycle, represented by a permutation $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ of cities, where π_k denotes the city visited at position k in the tour. In closed tour representations, the starting city may appear duplicated at the end of the sequence. However, the size of the problem n is defined as the number of distinct cities in the tour after removing the duplicated terminal node in the closed representation. The objective is to find a permutation that minimizes the total travel cost along the cycle, including the return to the starting city. Since the size of the search space increases factorially with n, exhaustive enumeration quickly becomes infeasible, motivating the use of heuristic and

metaheuristic approaches that explore neighborhoods defined over permutations to obtain high-quality solutions.

2.2 General Variable Neighborhood Search (GVNS)

GVNS operates over a predefined collection of neighborhood structures and combines a diversification phase, implemented through probabilistic shaking, with a deterministic intensification phase based on VND [13]. In VND, the available neighborhoods are explored in a fixed sequence: upon finding an improving move, the incumbent solution is updated and the search returns to the first neighborhood. If no improvement is found within the current neighborhood, the procedure moves to the next one. This continues until no further improvement is achievable under any available structure, at which point the solution is declared locally optimal. The VND routine is then embedded within an outer loop that alternates between shaking and local search, allowing the algorithm to escape local optima by periodically introducing controlled perturbations of increasing strength before resuming intensification. A complete description of the framework can be found in [12,13].

2.3 Adaptive Operator Selection (AOS)

AOS provides a control layer that regulates the choice of search operators according to their observed contribution during the search [10]. Rather than prescribing a specific learning strategy, it frames operator selection itself as a decision problem, delegating the underlying adaptation mechanism to the method in use. Within the machine learning in metaheuristics (ML-in-MH) paradigm, reinforcement learning has therefore been adopted as a practical means of implementing AOS in search processes that evolve over time [21].

In permutation-based problems such as the TSP, adaptive operator selection has been mainly investigated in population-oriented metaheuristics, including evolutionary algorithms and adaptive large neighborhood search [2,29], while trajectory-based single-solution approaches such as VNS and GVNS have received comparatively limited attention, typically relying on predefined operator schedules throughout the search process.

Reinforcement Learning. Reinforcement learning (RL) [31] is a method in which decisions are adapted over time based on feedback obtained from previous results. Rather than relying on predefined rules, the decision process progressively favors actions that have previously led to improvements, while discouraging those associated with ineffective moves. In the context of neighborhood-based metaheuristics, this allows the selection of local search operators to be guided by their observed contribution to solution quality during the search.

Each local search operator acts as a candidate action, and the quality change it induces in the solution serves as a natural performance signal. Because the usefulness of an operator at a given point in the search depends on the history of

moves already applied, the learning policy must account not only for immediate improvements but also for how operator transitions influence subsequent steps along the trajectory.

This framing places AOS within the broader class of sequential decision problems [10]. After each operator application, the observed outcome is used to update estimates of operator utility, which in turn shape the next selection. Repeated over many steps, this feedback loop gradually shifts effort toward operators that remain productive as the search evolves. Q-learning provides a principled way to close this loop [36], maintaining and updating value estimates for operator transitions without requiring any explicit model of the search landscape.

Q-Learning. Q-learning is an RL method that assigns a value to each possible action according to its expected outcome in a given state. These values are updated iteratively from the observed rewards, allowing the selection of actions that lead to improved results over time without requiring an explicit model of the underlying process.

The learning process is commonly formalized as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action set, P the transition probabilities, and R the reward function. At each decision step t, the agent observes the current state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$, receives a reward r_t, and transitions to a new state s_{t+1} [35].

Q-learning estimates the action–value function $Q(s, a)$, representing the expected discounted cumulative reward of selecting action a in state s. The estimates are updated iteratively according to the Bellman optimality:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a \in \mathcal{A}} Q(s_{t+1}, a) - Q(s_t, a_t) \right], \tag{1}$$

where $\alpha \in (0.0, 1.0]$ is the learning rate and $\gamma \in [0.0, 1.0)$ is the discount factor.

Under standard conditions (bounded rewards, Markov property, sufficient exploration), Q-learning converges with probability one to the optimal action–value function [31,35].

3 Proposed Algorithm

The proposed algorithm extends GVNS with RL–driven local search, using continuous credit assignment to adapt operator selection to the evolving search landscape. The Algorithm 1 receives as input an initialization method `InitMethod`, an iteration limit $I_{\max}$, a maximum neighborhood index $k_{\max}$, and a non–improvement limit $P_{\max}$. The initialization method controls how the initial tour is constructed and can assume one of two alternatives: a classical restricted candidate list nearest-neighbor heuristic (RCL) or a learning-based multi-agent reinforcement learning initializer (MARL), both described in detail in Sect. 3.1.

After initializing tour x and computing $c = f(x)$ (lines 1–2), the search proceeds in an outer loop of up to $I_{\max}$ iterations (line 4). Each iteration resets

$k = 1$ (line 7) and explores neighborhoods of increasing size in an inner loop (lines 8–20). At each level k, the incumbent is shaken (line 9) and then intensified by the RL local search (line 11). An improvement ($c'' < c$) triggers an update and resets $k = 1$ to restart from the first neighborhood (lines 12–16), otherwise k is incremented (line 18). The best tour found is returned after all iterations.

Algorithm 1 GVNS with RL-driven Local Search

Require: initialization method `InitMethod`, iteration limit $I_{\max}$, neighborhood limit $k_{\max}$, non-improvement limit $P_{\max}$

1: $x \leftarrow$ INITIALIZE(`InitMethod`)
2: $c \leftarrow f(x)$
3: $p \leftarrow 0$ ▷ consecutive non-improving iterations counter
4: **for** $i = 1$ to $I_{\max}$ **do**
5: **if** $p \geq P_{\max}$ **then break**
6: **end if**
7: $k \leftarrow 1$
8: **while** $k \leq k_{\max}$ **do**
9: $x' \leftarrow$ SHAKE(x, k)
10: $c' \leftarrow f(x')$
11: $(x'', c'') \leftarrow$ RLLOCALSEARCH(x', c')
12: **if** $c'' < c$ **then**
13: $x \leftarrow x''$
14: $c \leftarrow c''$
15: $k \leftarrow 1$ ▷ restart from the first neighborhood
16: $p \leftarrow 0$ ▷ reset non-improvement counter
17: **else**
18: $k \leftarrow k + 1$
19: **end if**
20: **end while**
21: $p \leftarrow p + 1$
22: **end for**
23: **return** x

3.1 Initialization

Two initialization procedures are employed. The first is a nearest–neighbor heuristic with an RCL mechanism [4], which introduces controlled variability in candidate city selection at each construction step. The second is a learning–based strategy inspired by MARL [1], where the tour is assembled incrementally using action–value estimates updated throughout construction.

The learning mechanism is adapted to the single-solution setting of GVNS rather than the population-based framework of [1]. Unlike MARL-based methods that update values only after complete tour evaluations, the proposed initializer employs a Q-learning formulation grounded in the Bellman optimality principle (Eq. 1), incorporating immediate and delayed rewards to account for the influence of early construction choices on subsequent search behavior.

3.2 Adaptive Shaking Strategy

In GVNS, diversification is achieved through a shaking phase that applies random perturbations to the incumbent solution, allowing transitions to neighboring regions of the search space [12]. In the proposed adaptive shaking formulation, (line 9 of Algorithm 1) and presented in Algorithm 2, the neighborhood index k is associated with the intensity of the perturbation by defining, at each shaking level, a set of admissible operators of increasing strength from which a move is randomly sampled.

Algorithm 2 Adaptive Shaking Procedure

Require: Current tour x, shaking strength k
Ensure: Perturbed tour x'
$n \leftarrow |x|$ ▷ number of cities in the tour representation
$\eta_k(n) \leftarrow$ size-dependent parameters ▷ segment lengths, exchange cardinalities
if $k \leq 1$ **then**
  $\mathcal{O}_k \leftarrow$ {Or-opt (small), Shuffle (mild)}
else if $k = 2$ **then**
  $\mathcal{O}_k \leftarrow$ {k-exchange (small), Cross-exchange (mild), Shuffle (medium)}
else if $k = 3$ **then**
  $\mathcal{O}_k \leftarrow$ {Double-bridge, k-exchange (medium), Cross-exchange (medium), Block recombination (medium)}
else
  $\mathcal{O}_k \leftarrow$ {Double-bridge, k-exchange (large), Cross-exchange (strong), Block recombination (strong), Shuffle (strong)}
end if
for a limited number of attempts **do**
  $o \sim \mathcal{U}(\mathcal{O}_k)$
  $x' \leftarrow o(x; \eta_k(n))$
  if $x' \neq x$ **then**
    return x'
  end if
end for
return x

The magnitude of the applied perturbation is controlled through operator parameters expressed as bounded functions of the size of the instance n. Consequently, the shaking phase operates on a set of established TSP perturbation operators, including segment relocations, edge exchanges, non-sequential perturbations, and block-based rearrangements, whose parameters scale with problem size.

In particular, Or–opt relocations [3], k–exchange moves [24], and cross–exchange operations [32] are applied using intensity–dependent parameterizations, while parameter-free operators such as double-bridge move [5,15] are employed at higher shaking levels. Additional variability is introduced through segment shuffling and block recombination mechanisms [28].

For segment-based operators, such as Or-opt [3] and cross-exchange [32], segment lengths are defined as bounded fractions of n, clipped between predefined minimum and maximum values. As a result, lower shaking levels modify short segments even in large instances, whereas higher levels operate on proportionally larger portions of the tour. Similarly, for exchange-based operators (e.g., k-exchange [24]), the number of exchanged nodes increases with n, while non-parametric operators such as double-bridge are applied without modification [5,15].

At each shake step, the neighborhood $\mathcal{N}_k(x)$ is implicitly defined by a set of perturbation operators $\mathcal{O}_k$ associated with the shaking level k (lines 3–11), whose parameters are scaled according to the size of the instance n (lines 1–2). The perturbation is then obtained by sampling an operator from $\mathcal{O}_k$ and applying it to the incumbent solution (line 14), with repeated attempts allowed to avoid trivial modifications (lines 12–18). The set of operators $\mathcal{O}_k$ is indexed by k to induce progressively stronger perturbations, ranging from mild structural changes to larger-scale tour rearrangements [5,15].

3.3 RL-Driven Local Search Procedure

The RLLocalSearch procedure implements the intensification phase of the proposed framework (line 11 of Algorithm 1). Starting from the shaken tour x', its objective is to exploit the local structure of the solution and produce an improved tour by sequentially applying local search operators. The selection of these operators is guided by a Q-learning mechanism combined with an ε-greedy policy, which allows adaptive control of the intensification process based on the observed effectiveness of operator transitions. The two components are independent: MARL (Sect. 3.1) operates only during construction with its own action-value estimates, while RL-GVNS maintains a separate Q-table over local search operator transitions.

Algorithm 3 describes the reinforcement learning–driven local search procedure used. The Q-table Q is initialized with zeros once before the outer GVNS loop (Algorithm 1) and is shared across all local search calls, allowing the agent to accumulate transition knowledge throughout the entire search. At the start of each call, the exploration rate is set to its current value ε, which decays after each call according to a multiplicative schedule (defined in Sect. 3.3), gradually shifting from exploration to exploitation.

The incumbent solution is set to the shaken tour x' and its objective value $c'' = f(x')$ is calculated (line 2). The set of operators is then initialized as $\mathcal{A} \leftarrow \mathcal{O}$ (line 4). The local search operates on a fixed set of neighborhood operators

$$\mathcal{O} = \left\{ \begin{array}{ll} \text{TwoOpt [9]}, & \text{OneMoveInsertion [11]}, \\ \text{ThreeOpt [6]}, & \text{DoubleBridge [15]} \end{array} \right\} \tag{2}$$

which differ in neighborhood size and structural impact. The initial operator is selected from $\mathcal{A}$ (line 4), and the local search proceeds while admissible operators remain (lines 6-19). At each iteration, the selected operator is applied to the

Algorithm 3 RL-driven Local Search

Require: Initial solution x', initial cost c', operator set $\mathcal{O}$, learning rate α, discount factor γ, initial exploration ε_0, minimum exploration $\varepsilon_{\min}$, exploration decay δ, max iterations $L_{\max}$, failure limit $F_{\max}$

Ensure: Improved solution x''

1: Q shared across calls (initialized once before GVNS loop); retrieve current ε
2: $x'' \leftarrow x'$, $c'' \leftarrow c'$
3: $\ell \leftarrow 0$, $f \leftarrow 0$ ▷ iteration and consecutive-failure counters
4: $\mathcal{A} \leftarrow \mathcal{O}$
5: $o \leftarrow \textsc{Select}(\mathcal{A}, Q, \varepsilon)$
6: **while** $\mathcal{A} \neq \emptyset$ **and** $\ell < L_{\max}$ **and** $f < F_{\max}$ **do**
7: $\quad \tilde{x} \leftarrow o(x'')$
8: $\quad \tilde{c} = f(\tilde{x})$
9: $\quad r \leftarrow \textsc{Reward}(c'', \tilde{c})$ ▷ reward design and Q-update
10: $\quad$ **if** $\tilde{c} < c''$ **then**
11: $\quad\quad x'' \leftarrow \tilde{x}$, $c'' \leftarrow \tilde{c}$
12: $\quad\quad \mathcal{A} \leftarrow \mathcal{O} \setminus \{o\}$ ▷ first-improvement acceptance
13: $\quad\quad f \leftarrow 0$ ▷ reset failure counter
14: $\quad$ **else**
15: $\quad\quad \mathcal{A} \leftarrow \mathcal{A} \setminus \{o\}$; $f \leftarrow f + 1$ ▷ operator pruning
16: $\quad$ **end if**
17: $\quad \ell \leftarrow \ell + 1$
18: $\quad$ Select next operator o from $\mathcal{A}$ using ε-greedy policy ▷ ε-greedy selection
19: **end while**
20: **return** x''

current solution, resulting in a candidate tour $\tilde{x}$ (line 7), which is subsequently evaluated to obtain its cost $\tilde{c}$ (line 8).

State–Action Model. Let $\mathcal{O} = \{o_1, \ldots, o_m\}$ denote the set of local search operators, and let $\tau : \mathcal{O} \to \{1, \ldots, m\}$ be an index map. The method maintains a Q-table $Q \in \mathbb{R}^{m \times m}$, where the state $s \in \{1, \ldots, m\}$ encodes the index of the previously applied operator and the action $a \in \{1, \ldots, m\}$ encodes the index of the next operator. Hence, $Q[s, a]$ estimates the utility of applying the operator a given that the previous operator was s. This operator-transition encoding is intentionally compact: it avoids the overhead of computing solution-quality indicators at every step, keeps the state space size fixed at m^2 regardless of instance size, and captures sequential operator dependencies that have been shown to influence local search effectiveness [10,36]. Richer state representations incorporating, e.g., current gap to best-known or stagnation counters, may improve learning precision but would increase both the state space and the tuning burden. This is left as a direction for future work.

Reward Design. After evaluating the candidate tour, the reward is calculated from the incumbent cost c'' and the candidate cost $\tilde{c}$ (line 8) as

$$r(c'', \tilde{c}) = \begin{cases} \dfrac{c'' - \tilde{c}}{\max(c'', 10^{-9})} & \text{if } \tilde{c} < c'', \\ -\lambda \cdot \dfrac{\tilde{c} - c''}{\max(c'', 10^{-9})} & \text{if } \tilde{c} > c'', \\ -0.01 & \text{if } \tilde{c} = c'', \end{cases} \tag{3}$$

where $\lambda \geq 1$ controls the asymmetry between positive and negative feedback: a cost reduction is rewarded in proportion to its relative magnitude, while a deterioration of equivalent size is penalized by λ. The small constant -0.01 assigned to neutral outcomes serves mainly as a tie-breaking signal, gently discouraging operators that produce no change without distorting the overall reward scale. After computing the reward, the Q-table is updated via the Bellman rule (Eq. 1), using the best estimated value for the successor state (line 9).

First-Improvement Acceptance and Operator Pruning. When the candidate tour improves on the incumbent ($\tilde{c} < c''$, line 10), the solution is immediately updated (line 11) and the admissible set is reset to $\mathcal{A} \leftarrow \mathcal{O} \setminus \{o\}$ (line 12), excluding the operator just applied so that the remaining operators can act on the improved solution. Conversely, if no improvement is found, the operator is dropped from $\mathcal{A}$ (line 15), avoiding repeated applications of a move that has already been tried without success in the current call.

ε-Greedy Operator Selection. After updating the admissible set, the next operator is selected from $\mathcal{A}$ according to an ε-greedy policy (line 18):

$$a \sim \begin{cases} \text{Uniform}(\mathcal{A}) & \text{with probability } \varepsilon, \\ \arg\max\limits_{a \in \mathcal{A}} Q[s, a] & \text{with probability } 1 - \varepsilon \end{cases} \tag{4}$$

which balances exploitation of high-utility operator transitions and exploration of alternative moves. The procedure ends when the admissible set becomes empty, returning the best solution found. After each call to the local search routine, the exploration rate is updated according to a multiplicative decay schedule,

$$\varepsilon \leftarrow \max(\varepsilon_{\min}, \varepsilon \cdot \delta) \tag{5}$$

with $\delta \in (0.0, 1.0)$, gradually shifting the policy from exploration to exploitation in successive calls.

4 Computational Experiments

The experimental design follows a 2×2 factorial structure, crossing two initializers (RCL or MARL) with two local search variants (fixed VND or RL-driven), yielding four configurations. This layout allows the contribution of each component to be assessed independently: varying the initializer while holding the local

search fixed measures the effect of the construction strategy, and varying the local search while holding the initializer fixed measures the effect of adaptive operator selection. All four configurations share the same adaptive shaking procedure (Algorithm 2). In standard GVNS, intensification follows a deterministic VND with the fixed operator order of Eq. 2. In RL-GVNS, it is replaced by the Q-learning-driven local search of Algorithm 3.

4.1 Methodology

All experiments were conducted on a single reference machine. The computational platform consisted of a MacBook Pro equipped with an Apple M3 Pro processor (11 cores) and 18 GB of RAM. All runs were executed in single-threaded mode, without parallelization. A representative subset of widely used TSPLIB (Traveling Salesman Problem Library) instances was selected to cover small, medium, and large-scale problem sizes, allowing a comprehensive evaluation of algorithmic performance. The corresponding instance characteristics and best known solution values reported in the literature are summarized in Table 2.

Table 2. TSPLIB instances used in this study, including problem size and best known solution values reported in the literature.

Instance	n	Best Known
bays29	29	2020
swiss42	42	1273
gr48	48	5046
berlin52	52	7542
kroA100	100	21282
kroB100	100	22141
ch150	150	6528
si175	175	21407
pr226	226	80369
pcb442	442	50778

For each algorithm–initialization pairing, defined by a specific search algorithm combined with a particular solution initialization strategy, 30 independent runs were conducted to account for stochastic variability. All methods used identical stopping criteria based on two conditions: a maximum number of iterations and a limit on consecutive non-improving iterations. For instances with at most 100 cities, the maximum number of iterations was set to 500 and the non-improvement limit to 200 consecutive iterations. For medium-sized instances ($100 < n < 200$) and larger instances ($n \geq 200$), these limits were reduced to 300 total iterations and 100 consecutive non-improving iterations, respectively. Performance was primarily evaluated on the final tour length, while solution stability across independent runs was also considered.

RL Parameters. Parameter values were set by grid search on three held-out instances (`berlin52`, `kroA100`, `pr226`), chosen to represent small, medium, and large problem scales. Candidate grids were built around ranges reported in the literature [1,36]: $\alpha \in \{0.10, 0.22, 0.30, 0.50\}$, $\gamma \in \{0.30, 0.50, 0.70, 0.80\}$, $\varepsilon_0 \in \{0.40, 0.60, 0.80\}$, $\delta \in \{0.95, 0.98, 0.995\}$, $\varepsilon_{\min} \in \{0.02, 0.05\}$, and $\lambda \in \{1.0, 1.2, 1.4, 2.0\}$. Rather than optimizing all parameters jointly, we followed the sequential protocol of [1], locking each group once selected before moving to the next. Every candidate setting was evaluated over 10 runs per calibration instance; the setting with the lowest mean gap was kept. A notable outcome was that the best values differ noticeably between small ($n \leq 100$) and larger instances ($n > 100$), which led to adopting separate parameter schedules for each regime. For the MARL initializer, the softmax temperature was raised above the value used in [1] to sharpen action selection, and the number of learning iterations was reduced for larger instances where additional iterations offered no further improvement. The resulting values are listed in Table 3.

Table 3. Parameters used in the proposed framework.

Parameter	Value(s)	Defined
Learning rate (α)	0.30 (small), 0.22 (large)	Algorithm 3, Eq. 1
Discount factor (γ)	0.50 (small), 0.80 (large)	Algorithm 3, Eq. 1
Initial exploration rate (ε_0)	0.60 (small), 0.40 (large)	Algorithm 3
Exploration decay (δ)	0.98 (small), 0.995 (large)	Algorithm 3
Minimum exploration ($\varepsilon_{\min}$)	0.05 (small), 0.02 (large)	Algorithm 3
Penalty scale (λ)	1.2 (small), 1.4 (large)	Eq. 3
Failure limit	10 consecutive failures	Algorithm 3
Max. local search iterations	300 (small), 600 (large)	Algorithm 3
Softmax temperature (β)	3	Sect. 3.1
MARL iterations	10,000 ($n < 100$), 5,000 ($n > 100$)	Sect. 3.1

Analysis. Table 4 reports the results obtained over 30 independent runs for each algorithm–initialization configuration. For each instance, *Best* denotes the best tour cost achieved, while *Max* and *Mean* correspond to the worst and average objective values, respectively. The *Gap (%)* represents the mean relative deviation from the best-known solution (BKS), calculated as:

$$\text{Gap}(\%) = \frac{1}{R} \sum_{i=1}^{R} \frac{c_i - \text{BKS}}{\text{BKS}} \times 100, \tag{6}$$

where c_i is the cost obtained in the i-th run and $R = 30$. The *BKSH* column reports the number of runs that reach the BKS. The *IterOpt* metric indicates the average number of iterations required to reach the best solution, and *Time (s)* denotes the mean execution time. For each instance, values highlighted in bold indicate the best-performing configuration for the corresponding metric.

Table 4. Performance results over 30 independent runs for each algorithmic configuration.

Instance	Alg.	Best	Max	Mean	Gap (%)	BKSH	IterOpt	Time (s)
bays29	(1)	**2020**	**2020**	**2020**	**0.00**	**30**	3.13	1.17
	(2)	**2020**	**2020**	**2020**	**0.00**	**30**	3.07	1.02
	(3)	**2020**	**2020**	**2020**	**0.00**	**30**	3.60	1.41
	(4)	**2020**	**2020**	**2020**	**0.00**	**30**	**2.60**	**0.91**
swiss42	(1)	**1273**	**1273**	**1273**	**0.00**	**30**	1.90	1.12
	(2)	**1273**	**1273**	**1273**	**0.00**	**30**	0.60	**0.10**
	(3)	**1273**	**1273**	**1273**	**0.00**	**30**	1.87	1.62
	(4)	**1273**	**1273**	**1273**	**0.00**	**30**	**0.50**	0.14
gr48	(1)	**5046**	**5046**	**5046**	**0.00**	**30**	9.47	11.73
	(2)	**5046**	**5046**	**5046**	**0.00**	**30**	**6.70**	**8.37**
	(3)	**5046**	**5046**	**5046**	**0.00**	**30**	11.93	17.80
	(4)	**5046**	**5046**	**5046**	**0.00**	**30**	9.57	12.27
berlin52	(1)	**7542**	**7542**	**7542**	**0.00**	**30**	3.07	3.38
	(2)	**7542**	**7542**	**7542**	**0.00**	**30**	1.47	**1.18**
	(3)	**7542**	**7542**	**7542**	**0.00**	**30**	3.23	5.25
	(4)	**7542**	**7542**	**7542**	**0.00**	**30**	**1.33**	1.56
kroA100	(1)	**21282**	**21282**	**21282**	**0.00**	**30**	15.87	73.07
	(2)	**21282**	**21282**	**21282**	**0.00**	**30**	8.20	**35.91**
	(3)	**21282**	**21282**	**21282**	**0.00**	**30**	17.37	120.76
	(4)	**21282**	**21282**	**21282**	**0.00**	**30**	**6.70**	48.44
kroB100	(1)	**22141**	22199	22162.27	0.10	19	**59**	683.12
	(2)	**22141**	22199	22150.67	0.04	25	75.16	**503.27**
	(3)	**22141**	22199	22144.87	0.02	28	71.54	1026.63
	(4)	**22141**	**22141**	**22141**	**0.00**	**30**	110.13	1212.83
ch150	(1)	**6528**	6564	6532.83	0.07	19	173.23	1967.42
	(2)	**6528**	6564	6534.70	0.10	21	125.00	1484.35
	(3)	**6528**	6566	6535.17	0.11	20	**113.74**	**1011.45**
	(4)	**6528**	**6549**	**6530.77**	**0.04**	**26**	131.33	1327.90
si175	(1)	**21407**	21421	21409.57	0.01	15	102.82	3387.24
	(2)	**21407**	21420	**21408.00**	**0.01**	17	180.00	3324.80
	(3)	**21407**	21427	21410.43	0.02	17	115.20	1204.38
	(4)	**21407**	**21419**	21410.89	**0.01**	**20**	171.00	**1067.30**
pr226	(1)	**80369**	80442	80381.00	**0.01**	27	83.00	5726.51
	(2)	**80369**	80442	80421.00	**0.01**	28	83.00	3512.51
	(3)	**80369**	80440	80371.37	**0.00**	**29**	**79.45**	2000.05
	(4)	**80369**	**80373**	**80369.13**	**0.00**	**29**	92.55	**1573.50**
pcb442	(1)	**50778**	51210	50950.00	0.10	18	135.00	5200.00
	(2)	**50778**	51150	50910.00	0.10	21	120.00	4800.00
	(3)	**50778**	51180	50930.00	0.07	23	110.00	6100.00
	(4)	**50778**	**50980**	**50850.00**	**0.06**	**27**	**95.00**	**4200.00**

The algorithmic configurations evaluated are: (1) RCL + GVNS, (2) MARL + GVNS, (3) RCL + RL-GVNS, and (4) MARL + RL-GVNS.

On small-scale instances, all four configurations reach the best-known solution in every run, and the differences between them are limited to convergence speed rather than solution quality. On `bays29` and `swiss42`, for example, all variants achieve a zero average gap and perfect BKSH rates, with only minor variation in iteration counts and runtimes. The picture changes as instance size increases. On

medium and large instances, the adaptive configurations produce more consistent results across independent runs, with lower mean gaps and higher best-known solution hit rates than the non-adaptive alternatives.

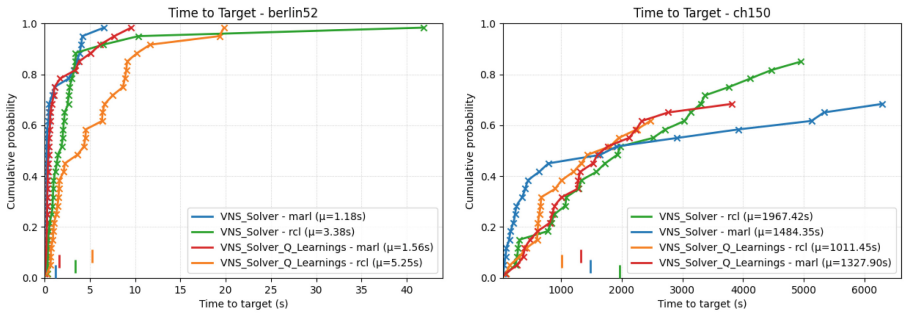
The runtime behavior is also noteworthy. Although the added complexity of MARL initialization and RL-guided local search might be expected to increase computation time, this is not consistently observed. For most instances, adaptive configurations reach high-quality solutions at least as fast as their non-adaptive counterparts, and in some cases faster. Better-informed operator sequencing appears to reduce unproductive search effort near high-quality regions of the landscape, where fixed operator orderings tend to lose effectiveness.

An exception is observed on `kroB100`, where the MARL + RL-GVNS configuration incurs a higher computational time (1 213 s) compared to its non-adaptive counterpart (503 s). However, this additional cost is accompanied by a marked improvement in robustness, as the adaptive variant is the only configuration to consistently reach the best-known solution in all 30 runs, yielding a zero average gap. In contrast, faster non-adaptive configurations exhibit residual gaps and lower BKSH rates. A similar pattern is observed on larger instances, such as `ch150` and `si175`, although with distinct trade-offs. For `ch150`, MARL-based configurations achieve lower average gaps (down to 0.04%) while also showing reduced intensification times and higher best-known solution hit rates, indicating that adaptive mechanisms improve both efficiency and solution quality. The same pattern is observed in `si175`.

As shown in Figs. 1a and 1b, all configurations eventually reach the best–known solution. However, the Empirical Cumulative Distribution Functions (ECDFs) reveal clear differences in the convergence dynamics. Learning–based configurations, particularly under MARL initialization, exhibit steeper early curves and a noticeable leftward shift, indicating a higher probability of reaching the target solution at earlier stages. In contrast, RCL–based variants show more gradual convergence, even when combined with learning. This effect becomes more pronounced in harder instances, where adaptive configurations converge faster and more consistently, in agreement with the lower average gaps and higher BKSH rates reported in Table 4.

To assess whether the observed performance differences are statistically significant, two complementary non–parametric analyses were performed following the framework of [8]. The Friedman test applied to mean tour costs across all 10 instances [8] produced $\chi^2 = 3.48$ ($p = 0.323$), which is not statistically significant. The five trivial instances (`bays29`, `swiss42`, `gr48`, `berlin52`, `kroA100`) yield tied ranks for all configurations, and restricting the analysis to the five hard instances produces the same statistic, as inconsistent behavior across instances (e.g., `si175`) prevents global significance. Average ranks favor the combined adaptive configuration: (4) MARL+RL–GVNS = 2.05, (2) MARL+GVNS = 2.55, (3) RCL+RL-GVNS = 2.65, and (1) RCL+GVNS = 2.75.

Pairwise Mann–Whitney U tests with Holm–Bonferroni correction [14] were applied to the five hard instances using run distributions derived from BKSH counts via two–point (BKS / max) imputation. Significant outcomes were found



(a) `berlin52` - All Combinations **(b)** `ch150` - All Combinations

Fig. 1. Time-to-target ECDF for representative TSPLIB instances.

for `kroB100` and `pcb442`. For `pcb442`, MARL+RL–GVNS (4) outperforms RCL+GVNS (1) with a medium–sized effect ($U = 603$, $p_{\text{adj}} = 0.0170$, $\hat{A}_{12} = 0.670$), confirming a statistically significant robustness advantage for the combined adaptive configuration. By contrast, the RCL+RL–GVNS (3) vs. MARL+RL–GVNS (4) comparison yields no significant difference ($U = 520$, $p_{\text{adj}} = 0.2546$, $\hat{A}_{12} = 0.578$), indicating that the adaptive local search alone may be insufficient without the MARL initializer on the hardest instances.

5 Conclusion

Classical GVNS applies local search operators in a fixed order regardless of how the search is evolving, a behavior that tends to become progressively less efficient as the algorithm approaches high-quality regions of the solution space. This work addressed that limitation by embedding a Q-learning mechanism directly into the intensification phase and pairing it with a MARL-based initialization that acts independently at the construction stage.

On small TSPLIB instances, all four configurations converge to the best-known solution in every run and the differences in runtime and iteration count are negligible. The more informative comparison comes from medium and large instances, where a static operator order begins to affect both solution quality and convergence consistency. The MARL + RL-GVNS configuration consistently produced lower average gaps and higher best-known solution hit rates, and in several cases it reached high-quality solutions faster than the non-adaptive variants. The `kroB100` instance illustrates a common pattern well. The adaptive configuration required roughly twice the wall-clock time of the fastest non-adaptive variant, yet it was the only one to reach the best-known solution in all 30 independent runs. The robustness advantage is substantial and repeatable across trials. Non-parametric statistical tests corroborate these observations: the Friedman test over all instances did not reach significance, largely because the five small instances produce tied ranks across configurations, but

pairwise tests on the harder instances confirmed statistically significant differences favoring MARL+RL-GVNS, with a medium-sized effect on `kroB100` and `pcb442`. Notably, the adaptive local search alone was not sufficient to produce a significant improvement on the hardest instances without the MARL initializer, suggesting that the two components are complementary rather than interchangeable.

The adaptive shaking strategy is shared by all four configurations and already improves upon a standard fixed-perturbation scheme. The RL components build on that foundation and do not replace it. The primary limitation of the current approach lies in the state representation, which encodes only the most recently applied operator. Extending it to include, for instance, the current gap relative to the best-known solution or a count of non-improving iterations would give the policy a richer basis for decisions, though at the cost of a larger Q-table and additional parameter tuning. Evaluating the framework on larger benchmark instances and adapting it to other neighborhood-based metaheuristics are the most natural directions for future work.

Acknowledgments. This work is partially supported by CAPES (Grant number #001) and FAPES (#2026-6NX7B).

Declaration of AI-Assisted Technologies. The use of the tool (Gemini - Google) was strictly limited to writing support and did not involve the generation of scientific content, results, or analysis. All AI-assisted text was subsequently reviewed and revised by the authors, who take full responsibility for the accuracy, integrity, and originality of the final work.

References

1. Alipour, M.M., Razavi, S.N.: A new multiagent reinforcement learning algorithm to solve the symmetric traveling salesman problem. Multiagent Grid Syst. **11**(2), 107–119 (2015)
2. Alipour, M.M., Razavi, S.N., Feizi Derakhshi, M.R., Balafar, M.A.: A hybrid algorithm using a genetic algorithm and multiagent reinforcement learning heuristic to solve the traveling salesman problem. Neural Comput. Appl. **30**(9), 2935–2951 (2018)
3. Babin, G., Deneault, S., Laporte, G.: Improvements to the or-opt heuristic for the symmetric travelling salesman problem. J. Oper. Res. Soc. **58**(3), 402–407 (2007)
4. Basu, S., Ghosh, D.: A review of the tabu search literature on traveling salesman problems. In: IIMA Working Paper WP2008-10-01. Indian Institute of Management Ahmedabad, Research and Publication Department (2008)
5. Bentley, J.L.: Fast algorithms for geometric traveling salesman problems. ORSA J. Comput. **4**(4), 387–411 (1992)
6. Blazinskas, A., Misevicius, A.: Combining 2-opt, 3-opt and 4-opt with k-swap-kick perturbations for the traveling salesman problem. In: Proceedings of the 17th International Conference on Information and Software Technologies (IT 2011), pp. 1–6. Kaunas, Lithuania (2011)
7. da Silva, R.F., Urrutia, S.: A general VNS heuristic for the traveling salesman problem with time windows. Discret. Optim. **7**(4), 203–211 (2010)

8. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
9. Feo, T.A., Resende, M.G.C.: Greedy randomized adaptive search procedures. J. Global Optim. **6**(2), 109–133 (1995)
10. Fialho, A., Da Costa, L., Schoenauer, M., Sebag, M.: Analyzing bandit-based adaptive operator selection mechanisms. Ann. Math. Artif. Intell. **60**(1), 25–64 (2010)
11. Glover, F., Rego, C.: Local search and metaheuristics for the traveling salesman problem. In: Gutin, G., Punnen, A. (eds.) The Traveling Salesman Problem and Its Variations, pp. 309–368. Kluwer Academic Publishers, Boston (2002)
12. Hansen, P., Mladenović, N., Moreno Pérez, J.A.: Variable neighbourhood search: methods and applications. 4OR **6**(4), 319–360 (2008)
13. Hansen, P., Mladenović, N., Todosijević, R., Hanafi, S.: Variable neighborhood search: basics and variants. Euro J. Comput. Optim. **5**(3), 423–454 (2017)
14. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**(2), 65–70 (1979)
15. Hoos, H.H., Stützle, T.: Stochastic local search. In: Handbook of Approximation Algorithms and Metaheuristics, pp. 297–307. Chapman and Hall/CRC (2018)
16. Hore, S., Chatterjee, A., Dewanji, A.: Improving variable neighborhood search to solve the traveling salesman problem. Appl. Soft Comput. **68**, 83–91 (2018)
17. Jeong, H.Y., Song, B.D.: Meta-learning-based adaptive operator selection for traveling salesman problem. Appl. Soft Comput. **185**, 113930 (2025)
18. Karakostas, P., Panoskaltsis, N., Mantalaris, A., Georgiadis, M.C.: Optimization of CAR T-cell therapies supply chains. Comput. Chem. Eng. **139**, 106913 (2020)
19. Karakostas, P., Sifaleras, A.: A double-adaptive general variable neighborhood search algorithm for the solution of the traveling salesman problem. Appl. Soft Comput. **121**, 108746 (2022)
20. Karakostas, P., Sifaleras, A., Georgiadis, M.C.: A general variable neighborhood search-based solution approach for the location-inventory-routing problem with distribution outsourcing. Comput. Chem. Eng. **126**, 263–279 (2019)
21. Karimi-Mamaghan, M., Mohammadi, M., Meyer, P., Karimi-Mamaghan, A.M., Talbi, E.G.: Machine learning at the service of meta-heuristics for solving combinatorial optimization problems: A state-of-the-art. Eur. J. Oper. Res. **296**(2), 393–422 (2022)
22. Kool, W., van Hoof, H., Welling, M.: Attention, learn to solve routing problems! In: International Conference on Learning Representations (2019)
23. Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., Shmoys, D.B.: The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization. Wiley, Chichester (1985)
24. Marx, D.: Searching the k-change neighborhood for tsp is W[1]-hard. Oper. Res. Lett. **36**(1), 31–36 (2008)
25. Menéndez, B., Bustillo, M., Pardo, E.G., Duarte, A.: General variable neighborhood search for the order batching and sequencing problem. Eur. J. Oper. Res. **263**(1), 82–93 (2017)
26. Mladenović, N., Hansen, P.: Variable neighborhood search. Comput. Oper. Res. **24**(11), 1097–1100 (1997), citeSeerX: 10.1.1.800.1797
27. Osman, I.H., Kelly, J.P.: Meta-heuristics: an overview. In: Osman, I.H., Kelly, J.P. (eds.) Meta-Heuristics: Theory and Applications, pp. 1–21. Kluwer Academic Publishers, Boston (1996)
28. Papalitsas, C., Karakostas, P., Andronikos, T.: A performance study of the impact of different perturbation methods on the efficiency of GVNS for solving TSP. Appl. Syst. Innov. **2**, 31 (2019)

29. Queiroz dos Santos, J.P., de Melo, J.D., Duarte Neto, A.D., Aloise, D.: Reactive search strategies using reinforcement learning, local search algorithms and variable neighborhood search. Expert Syst. Appl. **41**(10), 4939–4949 (2014)
30. Ruan, Y., Cai, W., Wang, J.: Combining reinforcement learning algorithm and genetic algorithm to solve the traveling salesman problem. J. Eng. **2024**, 2051–3305 (2024)
31. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, Adaptive Computation and Machine Learning. MIT Press, Cambridge (1998)
32. Taillard, E., Badeau, P., Gendreau, M., Guertin, F., Potvin, J.Y.: A tabu search heuristic for the vehicle routing problem with soft time windows. Transp. Sci. **31**(2), 170–186 (1997)
33. Todosijević, R., Mladenović, M., Hanafi, S., Mladenović, N., Crévits, I.: Adaptive general variable neighborhood search heuristics for solving the unit commitment problem. Int. J. Electr. Power Energy Syst. **78**, 873–883 (2016)
34. Venkatesh, P., Srivastava, G., Singh, A.: A general variable neighborhood search algorithm for the k-traveling salesman problem. Procedia Comput. Sci. **143**, 189–196 (2018), 8th International Conference on Advances in Computing & Communications (ICACC-2018)
35. Watkins, C.J.C.H.: Learning from Delayed Rewards. Ph.D. thesis, University of Cambridge, Cambridge (1989)
36. Wauters, T., Verbeeck, K., De Causmaecker, P., Vanden Berghe, G.: Boosting metaheuristic search using reinforcement learning. In: Talbi, E.G. (ed.) Hybrid Metaheuristics, pp. 433–452. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-30671-6_17

Computational Methods in Advanced Networking Technologies (CoMANT 2026)

When Heterogeneous Modalities Synchronize: Mitigating Desynchronization Attacks in Multimodal Online Federated Learning

Kainat Ibrar[1(✉)], Asim Ibrahim[2], and Francesco Palmieri[1]

[1] University of Salerno, Via Giovanni Paolo II, 132 - 84084 Fisciano, Salerno, Italy
{kibrar,fpalmieri}@unisa.it
[2] Department of Management Sciences, Bahria University Islamabad, Islamabad, Pakistan

Abstract. Multimodal online federated learning (MOFL) relies on accurate temporal alignment across heterogeneous data streams to enable effective cross-modal fusion. In this paper, we investigate temporal desynchronization as a practical adversarial attack surface in such systems. We show that even small temporal shifts between modalities can disrupt cross-modal alignment, leading to degraded model performance and increased attack effectiveness. Unlike traditional poisoning attacks, the proposed approach operates at the synchronization level, making it subtle and difficult to detect. We conduct experiments under both untargeted and label-conditioned (targeted) attack settings to comprehensively evaluate the impact of temporal misalignment. To mitigate this vulnerability, we propose a lightweight Cross-Client Consistency Scoring Mechanism (CCCSM) that reweights client updates based on their similarity during server-side aggregation. Experimental results on the UP-Fall dataset demonstrate that (i) the impact of the attack increases with the magnitude of temporal misalignment and (ii) the attack is more severe under targeted settings. For perturbation strength $k = 5$, the overall F1-score drops from 95.6% to 80.5%, while the attack success rate (ASR) increases up to 79.2% in targeted settings. The proposed CCCSM-based defense reduces ASR from 79.2% to 45.1% and improves overall F1 from 79.1% to 84.7% under targeted attacks with $k = 5$, while maintaining performance close to the clean baseline. These findings highlight the importance of synchronization in MOFL and the need for synchronization-aware defenses in real-world deployments.

Keywords: Online Federated Learning · Multimodal learning · Adversarial attacks · Cross-modal desynchronization attacks · Modality-Allignment attacks

1 Introduction

Federated learning (FL) has become a popular choice for modern intelligent systems, especially within safety-critical and distributed environments [15] where

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 371–386, 2027.
https://doi.org/10.1007/978-3-032-30521-3_23

data privacy and communication overhead constraints are of utmost importance for correct decision-making, such as in self-driving cars, industrial IoT, healthcare monitoring, and smart infrastructure. These systems use a variety of sensors, such as vision devices, inertial sensors, LiDAR, audio, and environmental sensors, to obtain additional information and improve decision-making [17] and are categorized as unimodal or multimodal FL systems, depending on the nature of the data possessed by the clients [7].

While traditional FL assumes static training rounds and synchronized updates, real-world deployments operate more and more frequently in Online Federated Learning (OFL) environments, in which clients continuously collect streaming data, train locally in sliding windows, and contribute model updates to a global server [20]. Moreover, in practical settings, the majority of the clients contain multimodal data, and this OFL's configuration introduces temporal drift, partial participation, and heterogeneous sampling rates across modalities, resulting in Multimodal Online Federated Learning (MOFL) systems heavily relying on accurate cross-modal synchronization to align heterogeneous sensor streams before fusion [20,21].

In the recent state-of-the-art (SOTA), synchronization mechanisms are assumed to be reliable and strong. The majority of current adversarial FL research focuses on backdoor attacks that target parameter updates, model replacement, and gradient poisoning [2,6], so that the temporal alignment layer remains an unexplored attack surface, especially in systems where synchronized heterogeneous modalities are crucial to fusion performance.

In this paper, we present a study that examines the exploitation of synchronization vulnerabilities in MOFL. We show that small and realistic temporal perturbations, such as timestamp shifts or window misalignments, can significantly degrade fusion quality and amplify adversarial effects. Unlike traditional poisoning attacks, these synchronization-based manipulations operate at the input alignment level, making them subtle, difficult to detect, and often indistinguishable from benign sensor delays or network jitter.

Specifically, we investigate temporal desynchronization attacks in MOFL, where adversarial clients introduce controlled temporal shifts between modalities to disrupt cross-modal alignment. By performing specific experiments on heterogeneous multimodal data, we demonstrate that even minor misalignments can lead to noticeable performance degradation and increased attack effectiveness under partial adversarial participation. To address this vulnerability, we propose a lightweight Cross-Client Consistency Scoring Mechanism (CCCSM), which reweights client updates based on their agreement during aggregation. Our results indicate that CCCSM effectively mitigates the influence of adversarial updates while maintaining performance on clean data.
Our work makes the following contributions:

1. We identify temporal synchronization in multimodal online federated learning as a potential adversarial attack surface.
2. We systematically study cross-modal desynchronization attacks and analyze their impact on multimodal fusion and model performance.

3. We demonstrate that temporal misalignment can induce both global performance degradation and class-specific prediction bias.
4. We propose and evaluate a lightweight consistency-based defense that improves robustness against desynchronization attacks.

Our research suggests that in streaming systems, robustness can collapse when modalities go out-of-sync, even under minor perturbations, highlighting the urgent need to reconsider synchronization assumptions in these systems.

The rest of the article is organized as follows: Sect. 2 reviews studies relevant to the proposed approach. Section 3 presents the system model, threat model, attack design, and CCCSM defense mechanism. Section 4 describes the dataset, preprocessing, baseline model, attack settings, and defense implementation. Section 4.6 discusses the results and key findings. Finally, Sect. 5 summarizes the study and outlines future directions.

2 Related Work

Multimodal learning systems rely on the fusion of heterogeneous data sources, where temporal synchronization is essential for constructing meaningful cross-modal representations. Prior work has shown that even minor temporal inconsistencies can significantly degrade fusion quality. For instance, sequence contrastive learning approaches employ dynamic time warping to align video and sensor streams, demonstrating the importance of precise temporal coordination [11]. Similarly, latent space alignment methods in healthcare explicitly enforce synchronization across modalities such as imaging and biosignals to ensure consistent feature representations [18]. Despite these advances, synchronization remains an active challenge rather than a fully resolved problem.

Multimodal datasets require careful synchronization between different modalities such as IMU signals, video streams, and textual annotations, highlighting that maintaining temporal alignment in real-world scenarios is non-trivial [10]. Similarly, O. Anigala et al. explore that temporal misalignment is a hidden data pathology that negatively affects model robustness and downstream performance [1]. These challenges become more pronounced in streaming environments, where data arrives continuously and must be aligned in real time.

In OFL, clients continuously generate and process streaming data, introducing additional complexity due to temporal drift, asynchronous updates, and heterogeneous sampling rates. For example, online spatio-temporal FL for traffic flow forecasting relies on dynamic temporal correlations and period-aware aggregation, implicitly assuming reliable temporal ordering [8]. Similarly, MOFL frameworks addressing missing modalities construct prototypes over synchronized temporal windows, again assuming consistent alignment across modalities [20]. In practical deployments such as precision agriculture systems, synchronization under network delays and clock drift remains an open challenge, with no standardized mechanisms to ensure temporal consistency [16]. These works highlight that OFL systems depend heavily on temporal integrity but do not consider adversarial manipulation of timestamps.

Recent advances in transformer-based and attention-driven multimodal fusion further rely on accurate temporal relationships. Attention mechanisms learn dependencies across modalities based on temporal ordering, as demonstrated in multimodal transformer architectures for coordinated attack detection [13]. Similarly, domain-driven threat detection and cross-modal behavioral prediction frameworks integrate temporal information across multiple data sources but operate at coarse temporal granularity and assume trustworthy timestamps [4,14]. As a result, fine-grained temporal perturbations remain largely undetected in such systems.

Robustness in FL has primarily been addressed through parameter-space defenses, including Byzantine-resilient aggregation and anomaly detection of model updates. Methods such as robust aggregation and trajectory-based anomaly detection aim to identify malicious updates based on their deviation from benign behavior [2,3]. However, these approaches operate on model updates rather than input data and therefore cannot detect attacks that corrupt the temporal alignment before model training.

In addition, graph-based and spatiotemporal models depend critically on correct temporal relationships for constructing meaningful structures. Federated graph neural networks and dynamic spatiotemporal models rely on timestamp-based dependencies, which can be invalidated by temporal perturbations [9,19]. Similarly, imputation-based methods for handling missing modalities depend on accurate temporal context, and their performance degrades when alignment is compromised [22].

Overall, existing literature assumes that temporal synchronization is reliable and focuses primarily on noise reduction, missing data, or parameter-level robustness. The adversarial implications of temporal misalignment, particularly in MOFL, remain largely unexplored. This gap motivates our work, which systematically investigates temporal desynchronization as an attack vector and proposes a cross-client consistency-based defense mechanism.

3 Methodology

We model a realistic MOFL pipeline with synchronized heterogeneous modalities and investigate how temporal desynchronization can be exploited as an adversarial mechanism. Figure 1 illustrates the overall workflow of the proposed framework, highlighting both benign and adversarial client pipelines. Unlike traditional attacks that explicitly manipulate labels or gradients, the proposed approach targets the synchronization process, leading not only to performance degradation but also to systematic misalignment in learned cross-modal representations. To address this, we propose a lightweight synchronization-consistency defense that improves robustness in streaming federated environments.

3.1 System Model

We consider an OFL setting with N distributed clients, where each client processes locally available multimodal data streams. In practice, we simulate the

OFL setting using a pre-collected multimodal dataset by treating each subject as a client and sequentially feeding temporally ordered data to mimic streaming behavior. Each client $i \in \{1, \ldots, N\}$ possesses a set of modalities:

$$\mathcal{M} = \{1, 2, \ldots, M\} \tag{1}$$

where each modality $m \in \mathcal{M}$ corresponds to a distinct sensor stream (e.g., wearable inertial sensors and camera-based observations). Due to heterogeneous sensing mechanisms, modalities operate at different sampling rates and temporal resolutions.

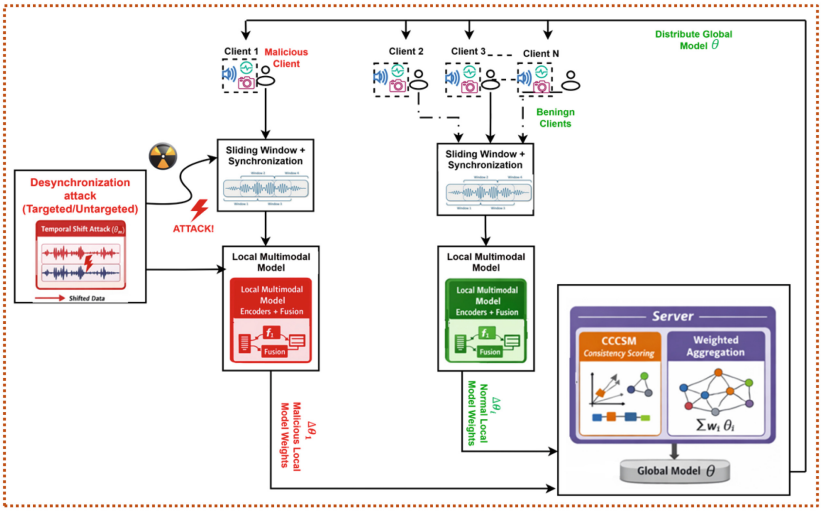


Fig. 1. The Overview of the Proposed Methodology

Temporal Windowing and Synchronization: To enable multimodal fusion, each client constructs aligned samples using a sliding window of size T. For modality m, the raw stream is denoted as

$$x_m(t) \tag{2}$$

A synchronization function $\mathcal{S}(\cdot)$ aligns modality-specific streams into a unified representation:

$$\tilde{x}_m^{(k)} = \mathcal{S}\big(x_m(t)\big), \quad t \in [t_k, t_k + T] \tag{3}$$

where k indexes the temporal window. This process involves resampling and timestamp alignment to ensure that multimodal observations correspond to the same underlying activity segment despite differing sampling rates.

Local Multimodal Model: Each client maintains modality-specific encoders:

$$z_m^{(k)} = f_m\big(\tilde{x}_m^{(k)}; \theta_m\big) \tag{4}$$

The modality embeddings are fused via a fusion function $\mathcal{F}(\cdot)$:

$$z^{(k)} = \mathcal{F}\big(z_1^{(k)}, z_2^{(k)}, \ldots, z_M^{(k)}\big) \tag{5}$$

The fused representation is passed to a task-specific head $h(\cdot)$ to produce predictions:

$$\hat{y}^{(k)} = h\big(z^{(k)}; \theta_h\big) \tag{6}$$

The local objective for client i is defined as:

$$\mathcal{L}_i = \mathbb{E}_{(x,y)}\big[\ell(\hat{y}, y)\big] \tag{7}$$

Online Federated Optimization: Clients update local models using temporally ordered data and periodically send model updates $\Delta\theta_i$ to the server. The server aggregates updates as:

$$\theta^{t+1} = \sum_{i \in \mathcal{C}_t} w_i \cdot \theta_i^t \tag{8}$$

where $\mathcal{C}_t$ denotes the set of participating clients at round t, and w_i represents aggregation weights.

3.2 Threat Model

We consider a subset of clients $\mathcal{A} \subset \mathcal{C}_t$ to be adversarial. The adversary has the following capabilities: (i) access to local multimodal data streams, (ii) ability to manipulate temporal alignment of modalities before fusion, and (iii) ability to participate intermittently in OFL. The adversary, however, does not control the server or the aggregation mechanism. The objective of the attacker is to degrade the global model performance, disrupt cross-modal fusion, induce systematic misclassification for specific activity patterns, and remain stealthy under standard aggregation rules.

3.3 Cross-Modal Desynchronization Attack

We introduce a synchronization-based attack that perturbs temporal alignment between modalities.

Attack Formulation: After synchronization, for a target modality m, the adversary applies a temporal shift:

$$x_m(t) \rightarrow x_m(t + \delta_m) \tag{9}$$

where δ_m is a small time offset relative to the sampling rate. The synchronized representation becomes:

$$\tilde{x}_m^{(k)} = \mathcal{S}\big(x_m(t + \delta_m)\big) \tag{10}$$

while other modalities remain unchanged.

Attack Impact: Fig. 2 illustrates the effect of temporal alignment on multimodal fusion, comparing synchronized and desynchronized modality streams. It shows that even small temporal shifts (δ_m) lead to misaligned signals, resulting in inconsistent fusion and degraded representation learning.

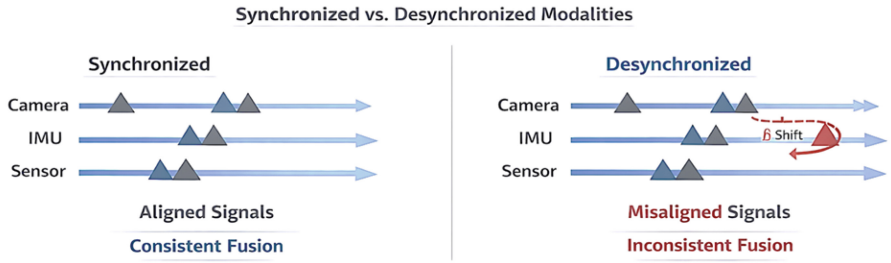


Fig. 2. The Comparison of Synchronized and Desynchronized Multimodal Signals

The intuition behind this attack lies in the assumption that multimodal fusion requires modality-specific encodings to represent the same underlying temporal event. Formally:

$$z_m^{(k)} \approx z_{m'}^{(k)}, \quad \forall m, m' \in \mathcal{M} \tag{11}$$

Desynchronization violates this assumption and illustrates that (a) embeddings correspond to different temporal contexts, (b) fusion introduces inconsistent cross-modal features and (c) the model learns spurious or incorrect relationships between modalities.

Although the perturbation is applied at the input level, its effect propagates through the learning process. In particular, it is executed in these two settings:

– **Global degradation:** In this setting, the temporal shift is applied to all samples at adversarial clients. Misaligned multimodal inputs introduce noise in representation learning, reducing overall model accuracy.

– **Implicit targeted effect:** In this setting, we apply temporal shifts selectively to samples belonging to a specific activity class; the model is exposed to consistently misaligned patterns for that class. Over multiple OFL rounds, this results in systematic confusion and biased predictions for those activities.

Moreover, the OFL setting itself is inherently susceptible to amplification of desynchronization effects due to continuous exposure to temporally ordered data, temporal drift in local data distributions, and asynchronous and partial client participation. Overall, our proposed attack exhibits the following characteristics:

1. Even small values of δ_m (e.g., on the order of milliseconds) are sufficient to significantly degrade model performance.
2. The perturbation closely resembles natural sensor delays, clock drift, or network-induced latency. As a result, it is difficult to distinguish adversarial manipulation from benign temporal inconsistencies.
3. It targets the synchronization layer rather than the model parameters or labels, thereby disrupting cross-modal representation learning.
4. It does not depend on a specific fusion strategy or model architecture. It can be applied to early, late, or attention-based fusion mechanisms, making it broadly applicable across different MOFL systems.

3.4 Cross-Client Consistency Scoring Mechanism (CCCSM)

The proposed Cross-Client Consistency Scoring Mechanism (CCCSM) is based on the observation that in a well-synchronized multimodal setting, benign client updates tend to exhibit consistent optimization directions. In contrast, desynchronization disrupts cross-modal alignment at the representation level, leading to inconsistent and noisy model updates.

Server-Side Consistency Scoring: CCCSM operates entirely at the server side and does not require access to raw data. For each participating client i, the server computes a consistency score based on the similarity [23] between its update $\Delta\theta_i$ and those of other clients in the same round:

$$c_i = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \cos(\Delta\theta_i, \Delta\theta_j) \tag{12}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity.

Clients Weighting:

The computed consistency scores are used to reweight client contributions during aggregation by using the following Eq. (13)

$$w_i = \frac{c_i}{\sum_{j \in \mathcal{C}_t} c_j} \tag{13}$$

A lower consistency score indicates potential desynchronization-induced deviation. Therefore, clients with lower consistency scores are down-weighted, reducing their influence on the global model.

The key properties of our proposed defense are that it is model-agnostic, relies only on lightweight consistency scoring over client updates, and does not require access to labeled validation data or additional local training, thereby incurring low computational overhead and making it suitable for wearable and IoT-based MOFL deployments with limited memory, battery, and processing resources.

However, the framework has certain limitations as well. In particular, CCCSM assumes that benign clients constitute the dominant portion of participating updates in each round. Under coordinated malicious-majority settings, update similarity alone may be insufficient for reliable discrimination.

4 Experimental Setup

4.1 Dataset

We conduct our evaluation using the UP-Fall dataset [12], a multimodal human activity recognition dataset designed for fall detection and daily activity monitoring. The dataset consists of 11 activities, including six daily activities and five types of falls, performed by 17 subjects under controlled conditions.

The dataset provides synchronized multimodal recordings, including wearable sensor signals (accelerometer and gyroscope) and vision data captured from multiple cameras. These modalities exhibit different temporal resolutions and sensing characteristics, making the dataset suitable for studying cross-modal synchronization in MFL.

4.2 Preprocessing and Feature Construction

Following the feature construction pipeline provided with the UP-Fall dataset [12], the raw multimodal streams are segmented using sliding windows of duration $T \in \{1, 2, 3\}$ seconds with 50% overlap. Unless otherwise stated, we report results for $T = 2$ seconds as a balanced configuration between temporal resolution and contextual information.

For the sensor modality, handcrafted features are extracted from each windowed segment. Specifically, 18 features per signal are used, including temporal statistics (e.g., mean, standard deviation, quartiles) and frequency-domain features (e.g., spectral energy and entropy). Given the multiple sensor channels, this results in a high-dimensional feature representation per window.

For the vision modality, we adopt the Horn and Schunck optical-flow-based representation [5,12]. Motion between consecutive frames is computed, transformed into magnitude representations, resized to a fixed spatial resolution, and aggregated within each window to obtain a compact feature vector.

The resulting modality-specific features are normalized and used as inputs to the multimodal learning model.

4.3 Baseline Implementation Model

We employ a lightweight multimodal architecture consisting of modality-specific encoders and a fusion module. Each modality is processed using a dedicated multilayer perceptron (MLP) encoder to obtain latent embeddings. The embeddings are fused via concatenation and passed to a classification head for activity prediction. This design enables effective integration of heterogeneous modalities while maintaining computational efficiency, which is important in FL environments.

We simulate an MOFL environment by treating each subject as an individual client, resulting in $N = 17$ clients. Although the dataset is pre-collected, we emulate streaming behavior by feeding temporally ordered windowed samples sequentially to each client. Training is performed over 100 communication rounds. At each round, a subset of clients (10%) is randomly selected to participate in training. Each selected client updates its local model using newly available data segments and sends model updates to the server for aggregation.

To simulate realistic non-IID data distributions, we partition the data using a Dirichlet distribution over activity labels. We set the concentration parameter α, which governs the degree of data heterogeneity across clients, to 0.3, representing a moderately high level of non-IID distribution. Moreover, the local model is trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The hidden dimension of the MLP encoders is set to 128 with ReLU activation. The proposed CCCSM defense is applied at the server side during aggregation by reweighting client updates based on consistency scores.

4.4 Attack Configuration

The attack is realized by perturbing the temporal alignment of a selected modality at the raw signal level before the synchronization and windowing process. The detailed procedure of the proposed cross-modal desynchronization attack is presented in Algorithm 1.

For a given modality m, the adversarial transformation is defined as Eq. (9). Rather than defining δ_m as an absolute time value, we express it relative to the sampling interval of the modality:

$$\delta_m = k \cdot \Delta t_m \tag{14}$$

Here, Δt_m denotes the sampling interval and k is the number of shifted time steps. In our experiments, we consider small values of $k \in \{1, 2, 3, 5\}$. In our dataset, wearable sensor signals are sampled at approximately 50 Hz, corresponding to a sampling interval of $\Delta t_m \approx 20$ ms. Therefore, these values of k result in temporal offsets in the range of approximately 20–100 ms. For the vision modality, which is captured at a fixed frame rate, the perturbation corresponds to shifting the sequence by a small number of frames, leading to minor frame-level misalignment across modalities.

This perturbation is applied prior to the synchronization function $\mathcal{S}(\cdot)$, ensuring that the resulting multimodal windows contain temporally inconsistent information across modalities.

The attack is executed locally at adversarial clients during each communication round. For each selected client, the temporal perturbation is applied consistently across its local training data according to the attack configuration. After perturbation, the client performs standard local training and sends model updates to the server. Consequently, the adversarial effect is embedded implicitly in the learned representations and propagates through the federated aggregation process.

Untargeted and Label-Conditioned Settings To comprehensively evaluate the impact of desynchronization, we consider two attack settings:

Untargeted Desynchronization: In this setting, the temporal shift is applied to all samples at adversarial clients. The primary effect is a global degradation in model performance due to disrupted cross-modal alignment. The model receives inconsistent multimodal inputs, leading to noisy representations and reduced predictive accuracy.

Label-Conditioned Desynchronization: In this setting, the perturbation is applied selectively to samples belonging to a specific source class c_s. Only samples with label $y = c_s$ are temporally shifted, while all other samples remain unchanged. Over multiple communication rounds, this selective exposure introduces a systematic bias in the learned representations for class c_s. The model associates inconsistent multimodal patterns with this class, leading to increased confusion with other activities. This behavior resembles a targeted adversarial effect, even though no explicit trigger or label manipulation is introduced.

4.5 Defense Implementation

We apply the proposed CCCSM during server-side aggregation at each communication round. Upon receiving client updates $\{\Delta\theta_i\}_{i\in\mathcal{C}_t}$, the server first flattens each update into a vector form and computes pairwise cosine similarity between updates [23]. The consistency score for each client i calculated as Eq. (12), is defined as the average similarity with respect to all participating clients:

To ensure stable aggregation, the scores are normalized using a softmax function:

$$w_i = \frac{\exp(c_i)}{\sum_{j\in\mathcal{C}_t}\exp(c_j)} \tag{15}$$

These normalized weights are used to reweight client contributions during aggregation, thereby reducing the influence of inconsistent or adversarial updates.

Algorithm 1 Cross-Modal Desynchronization Attack in MOFL

Require: Participating clients $\mathcal{C}_t$, adversarial subset $\mathcal{A}_t$, target modality m, shift step k, sampling interval Δt_m, optional source class c_s
Require: Window size T, synchronization function $\mathcal{S}(\cdot)$, global model θ^t
Compute temporal shift $\delta_m \leftarrow k \cdot \Delta t_m$
for all client $i \in \mathcal{C}_t$ **do**
  Initialize local model with global parameters θ^t
  for all local samples (x, y) at client i **do**
    if $i \in \mathcal{A}_t$ **then**
      if c_s is not specified **or** $y = c_s$ **then**
        Apply temporal shift to modality m:

$$x_m(t) \leftarrow x_m(t + \delta_m)$$

      end if
    end if
    Construct synchronized windowed sample using $\mathcal{S}(\cdot)$ over interval $[t_k, t_k + T]$
  end for
  Perform local training on the resulting multimodal windows
  Compute local update $\Delta\theta_i^t$
  Send $\Delta\theta_i^t$ to server
end for
Server aggregates updates to obtain θ^{t+1}

In implementation, the mechanism is applied at every communication round without introducing additional communication overhead, as it operates solely on received model updates.

4.6 Results and Discussion

We evaluate the system under three experimental settings. First is the *no attack* scenario, where standard MOFL is performed without any adversarial intervention, serving as a baseline for comparison. Second is the *desynchronization attack* setting, in which a subset of adversarial clients introduces temporal shifts in one modality during local training. Finally, we consider the *attack with CCCSM* setting, and apply proposed defense mechanism at the server during aggregation to mitigate the impact of such adversarial updates.

Performance is evaluated using attack-oriented metrics widely adopted in the literature [6], including the attack success rate (ASR), F1-score on clean samples (Normal F1), F1-score on perturbed samples (Attack F1), and overall F1-score.

Table 1 presents a detailed evaluation of the proposed desynchronization attack under varying temporal shift strengths and attack configurations.

First, we observe that even small temporal shifts ($k = 1$) lead to noticeable degradation in performance, confirming the sensitivity of multimodal fusion to temporal alignment. As the shift magnitude increases, the attack success rate consistently rises, while both Attack F1 and Overall F1 decline. This trend indi-

Table 1. Attack-oriented evaluation under untargeted and label-conditioned desynchronization for different shift strengths k. Results are averaged across clients on the UP-Fall dataset.

Setting	k	ASR (%)	Normal F1 (%)	Attack F1 (%)	Overall F1 (%)
No Attack (Baseline)					
Baseline (MOFL)	–	0.0	95.6	–	95.6
Without Defense					
Untargeted	1	40.9	91.3	87.2	89.1
	3	56.8	89.4	80.9	85.0
	5	68.9	87.1	75.4	80.5
Targeted	1	52.6	90.8	85.1	87.6
	3	67.9	88.8	79.6	84.2
	5	79.2	86.5	73.0	79.1
With CCCSM Defense					
Untargeted + CCCSM	1	19.4	92.1	90.0	91.1
	3	28.1	90.9	86.4	89.0
	5	39.8	89.3	82.2	85.6
Targeted + CCCSM	1	25.2	91.6	88.7	90.2
	3	33.4	90.2	85.8	88.7
	5	45.1	88.8	81.0	84.7

cates that stronger temporal misalignment introduces progressively severe inconsistencies in the learned cross-modal representations.

We also note that the label-conditioned (targeted) setting consistently exhibits higher ASR compared to the untargeted case, along with slightly lower Attack F1 and Overall F1 values. This suggests that selectively applying temporal perturbations to specific activity classes induces stronger and more focused misclassification behavior, effectively acting as a targeted adversarial mechanism even in the absence of explicit trigger patterns.

Additionally, our results demonstrate that the proposed CCCSM defense significantly reduces ASR across all values of k and improves both Attack F1 and Overall F1. Notably, the improvement is more pronounced for smaller perturbations, while the recovery becomes less effective as the shift magnitude increases. This behavior indicates that consistency-based weighting effectively suppresses anomalous client updates and stabilizes the learning process, particularly under mild to moderate attack conditions.

However, it is important to mention that the performance does not fully recover to the no-attack baseline, especially for larger values of k. This suggests that temporal desynchronization introduces persistent distortions in cross-modal representations that cannot be entirely eliminated through aggregation alone.

To provide further insight into the effect of temporal desynchronization, Fig. 3 illustrates the variation of Overall F1 with respect to the shift strength k. From the plot, it is evident that model performance degrades consistently as k increases, highlighting the sensitivity of multimodal fusion to temporal misalignment. This degradation is more pronounced in the targeted setting, where we apply temporal perturbations on selected labeled classes.

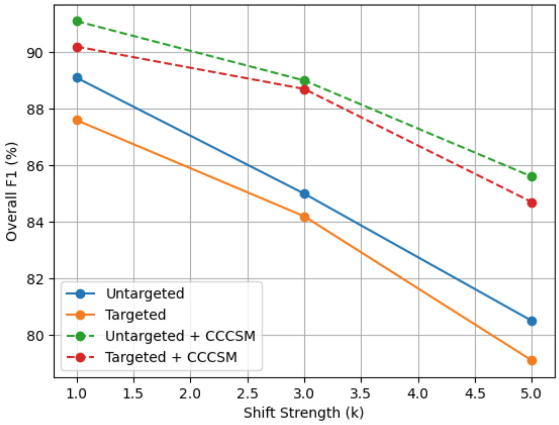


Fig. 3. Impact of temporal shift strength k on model performance.

Overall, the results demonstrate that temporal desynchronization constitutes a practical and effective attack vector in MOFL, while the proposed CCCSM defense provides a lightweight and effective mitigation strategy.

5 Conclusion and Future Work

In this paper, we investigated temporal desynchronization as a practical adversarial threat in MOFL and showed that even small temporal shifts can significantly degrade performance and induce systematic errors. We further demonstrated that the impact of the attack increases with the magnitude of misalignment and is more pronounced in targeted settings. To address this threat, we proposed CCCSM, a lightweight defense that effectively reduces the ASR and improves robustness without requiring additional data. Furthermore, since the proposed approach is modality-agnostic, it is not limited to the specific sensor–image modalities considered in this study and can be extended to other multimodal datasets and application domains.

However, our results also show that CCCSM does not fully recover performance under stronger perturbations, motivating the development of more adaptive synchronization-aware defenses. Moreover, the evaluation is conducted using a pre-recorded wearable-sensing dataset; streaming behavior is mimicked in a controlled and reproducible setting. Therefore, future work includes evaluating CCCSM on larger and more diverse datasets, comparing it against established FL defenses, and assessing its efficiency and robustness in real-world asynchronous federated environments.

Disclosure of Interests. There is no conflict of interest.

References

1. Anigala, O., Okolo, J.N., Son, J., Pack, C.: Mm-score: a two-factor framework for auditing multimodal dataset quality. In: Proceedings of the International Conference on Research in Adaptive and Convergent Systems, pp. 1–8 (2025)
2. Batur Şahin, C.: Securing UAV swarms with vision transformers: a byzantine-robust federated learning framework for cross-modal intrusion detection. Drones **10**(2), 125 (2026)
3. Ding, Z., et al.: Identifying alternately poisoning attacks in federated learning online using trajectory anomaly detection method. Sci. Rep. **14**(1), 20269 (2024)
4. Gu, Z., Guo, J., Xu, J., Sun, Y., Liang, W.: Domain knowledge-driven method for threat source detection and localization in the power internet of things. Electronics **14**(13), 2725 (2025)
5. Horn, B.K., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)
6. Lai, Y., Wu, L., Lin, H., Liu, J.: Bad-MFL: a cross-modality bi-trigger backdoor attack against multi-modal federated learning. IEEE Internet Things J. (2025)
7. Li, L., Fan, Y., Tse, M., Lin, K.Y.: A review of applications in federated learning. Comput. Ind. Eng. **149**, 106854 (2020)
8. Liu, Q., Sun, S., Liu, M., Wang, Y., Gao, B.: Online spatio-temporal correlation-based federated learning for traffic flow forecasting. IEEE Trans. Intell. Transp. Syst. **25**(10), 13027–13039 (2024)
9. Liu, R., Xing, P., Deng, Z., Li, A., Guan, C., Yu, H.: Federated graph neural networks: Overview, techniques, and challenges. IEEE Trans. Neural Networks Learn. Syst. (2024)
10. Ma, J., et al.: T3Set: a multimodal dataset with targeted suggestions for LLM-based virtual coach in table tennis training. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, vol. 2. pp. 5686–5697 (2025)
11. Ma, Y., He, H., Wang, X., Wu, G.: Sequence contrastive learning for multi-modal alignment in heavy plate rolling. In: 2025 7th International Conference on Industrial Artificial Intelligence (IAI), pp. 1–6. IEEE (2025)
12. Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., Peñafort-Asturiano, C.: Up-Fall detection dataset: a multimodal approach. Sensors **19**(9), 1988 (2019)
13. Nooraien, M.A., Kumar, P.D., Priya, M.B.L., Nadeem, P., et al.: A multimodal transformer architecture for detecting coordinated network attacks. In: 2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN), pp. 992–997. IEEE (2025)
14. Pu, X., Ren, Y., Liang, J.: ACMBP: an adaptive graph-based meta-learning framework for cross-modal behavior prediction from IoT and social media streams. Informatica **50**(9) (2026)
15. Quan, M.K., et al.: Federated learning for cyber physical systems: a comprehensive survey. IEEE Commun. Surv. Tutorials (2025)
16. Rao, S., Neethirajan, S.: Computational architectures for precision dairy nutrition digital twins: a technical review and implementation framework. Sensors **25**(16), 4899 (2025)
17. Sahoo, S.: Sensor fusion and virtual sensor design for enhanced multi-sensor data accuracy in autonomous systems. Int. J. Smart Sustain. Intell. Comput. **1**(2), 21–39 (2024)

18. Shrivastava, A., Kholmatovna, U.N., Kannimuthu, S., MohamadAbbas, H., Thakur, D., Priyanka, T.P.: Cross-modal neural networks for real-time multi-sensory information fusion in smart healthcare. In: 2025 International Conference on Innovations and Emerging Technologies In AI & Communication Systems (IETACS), pp. 1567–1570. IEEE (2025)
19. Tan, C., Hu, J., Li, J., Miao, M., Hu, W., Wang, S.: DTS-MixNet: dynamic spatiotemporal graph mixed network for anomaly detection in multivariate time series. Big Data Cogn. Comput. **9**(10), 245 (2025)
20. Wang, H., Liu, X., Zhong, X., Chen, L., Liu, F., Zhang, W.: Multimodal online federated learning with modality missing in internet of things. IEEE Trans. Mobile Comput. (2025)
21. Wang, H., Zhong, X., Liu, K., Liu, F., Zhang, W.: Denoising and adaptive online vertical federated learning for sequential multi-sensor data in IIoT. IEEE Trans. Mobile Comput. (2025)
22. Wang, J., et al.: Patch: a plug-in framework of non-blocking inference for distributed multimodal system. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 7(3), pp. 1–24 (2023)
23. Xia, P., Zhang, L., Li, F.: Learning similarity with cosine similarity ensemble. Inf. Sci. **307**, 39–52 (2015)

A Fractional Epidemic Model for Multi-malware Attacks in Wireless Sensor Networks

Eslam Farsimadan[1(✉)], Leila Moradi[1], Marcello Trovati[2], and Francesco Palmieri[1]

[1] Department of Computer Science, University of Salerno, 84084 Fisciano, Italy
{efarsimadan,lmoradi,fpalmieri}@unisa.it
[2] School of Business, University of Lancashire, Preston, Lancashire, UK
mtrovati@lancashire.ac.uk

Abstract. Nowadays, Wireless Sensor Network (WSN) technology is widely used in many applications requiring distributed monitoring and control capabilities. When large WSN infrastructures are involved, the diffusion of malware among nodes becomes a critical issue, requiring a clear understanding of propagation dynamics to guarantee timely countermeasures. By considering that fractional-order models demonstrated to be particularly effective in forecasting the dynamics of complex real-life processes, this study improves the Susceptible-Exposed1-Exposed2-Infected-Recovered (SE_1E_2IR) model, derived from the traditional SEIR epidemic theory and used for the early detection of multi-malware activities in WSNs. In particular, a fractional SE_1E_2IR epidemic model has been formulated to explain the propagation of multiple malware infections according to a more versatile and realistic view. The model is structured upon a system of fractional differential equations. The basic reproduction number is also determined as a crucial factor for characterizing malware propagation within the network. The model's free and endemic equilibrium points are calculated, and the stability of the equilibrium points is investigated to find requirements for preventing continuous malware propagation in WSNs. To complement the theoretical analyses, MATLAB is used to perform numerical simulations.

Keywords: Wireless sensor networks · Malware propagation · Epidemic models · Basic reproductive number · Stability analysis

1 Introduction

The mechanism of malware propagation in scale-free networked systems is very similar and somewhat comparable to that of biological virus infections in human populations. The propagation of biological viruses in populations has been extensively studied through epidemic theories, resulting in the development of several theoretical models characterizing this process [2]. Consequently, mathematical

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 387–405, 2027.
https://doi.org/10.1007/978-3-032-30521-3_24

models that describe the spreading and propagation dynamics of infectious diseases in the population and malware in the network can assume a significant role in both theoretical and practical contexts. Thus, by developing and investigating these models, one can better comprehend the processes, conditions, and attributes of malware as it moves throughout the network. For instance, the Susceptible-Infected-Removed (SIR) model and the Susceptible-Infected-Susceptible (SIS) model are two traditional models derived from epidemic theory, frequently used to characterize the dynamics of worm propagation on the Internet. Nevertheless, limited effort has been devoted to malware propagation models that are specific for WSNs. In several recent studies, researchers have used epidemic models to study how malware spreads in WSNs [18,22].

The majority of models that have been suggested to study the spread of malware in these networks rely on Ordinary Differential Equations (ODEs), which are an ineffective tool for describing the behaviors of a complicated networked system. To address these weaknesses, we consider the use of Fractional Differential Equations (FDEs) as an effective technique for modeling malware propagation in WSNs. Indeed, FDEs possess the capacity to provide a more precise characterization of systems displaying non-smooth or discontinuous behavior due to their capability of capturing the memory effect. Researchers have recently expressed their interest in employing FDEs to model malware propagation in network systems, for instance, in [19,21].

Motivated by the issues mentioned previously, we develop and investigate a new mathematical model, starting from the existing Susceptible-Exposed1-Exposed2-Infected-Recovered (SE_1E_2IR) multi-malware epidemic model, originally based on ODEs [1], that explains malware propagation in WSNs by using FDEs, in order to provide more accurate results and avoid the memory limitations characterizing integer order formulations. More precisely, Caputo fractional derivatives have been used for achieving unrestricted memory and increased flexibility in modeling intricate dynamics because they better capture the memory effect and long-range dependence characteristics of complex systems. Therefore, they are more appropriate for modeling and analyzing real-world phenomena, such as fractal behavior, anomalous diffusion, and non-local interactions. Implementing fractional-order models also shows great potential in precisely depicting the complex behavior of multiple malware infections simultaneously spreading throughout a WSN infrastructure. Therefore, a fractional-order $0 \leq \alpha < 1$ is used to characterize the dimensional oddities characterizing multi-malware propagation and support the detection of malware existence in WSNs. The Fractional Adam-Bashforth-Moulton (FABM) technique is then employed as an appropriate corrective technique to provide hints on the identification or elimination of the malware infection. The stability conditions are also investigated. Simulation results show that the fractional-order model outperforms the integer-order propagation one.

The rest of this paper is organized as follows: Sect. 2 reports some related work. In Sects. 3, the background, description, and formulation of the proposed model are presented. Section 4 discusses the basic properties of the model. The

system's stability, its relevant and important theorems with their proof are investigated in 5. The simulation outcomes of the modified model and the discussion of the achieved results are presented in Sect. 6. Finally, Sect. 7 presents the conclusion.

2 Related Work

To effectively mitigate and control malware propagation in WSNs, it is necessary to assess and forecast the malware behavior by leveraging reliable models. Recent studies on malware propagation on networked systems, such as WSNs, are based on epidemic theories, which, in turn, rely on mathematical models where differential equations play an essential role.

Verma and Gupta [20] introduced a Vaccination-Susceptible-Carrier-Infectious-Quarantine-Recovered-Susceptible (VSCIQRS) model to mitigate malware propagation in WSNs. The suggested method is employed to prevent malicious activities within the networks. The vaccination technique was employed to mitigate malware propagation in WSNs and could potentially maximize the system's anti-malware abilities. Recently, Dong et al. [7] developed a fractional network-based propagation model featuring a saturated treatment function and developed a fuzzy controller to examine infection management within complex heterogeneous networks. An overview of various malware propagation models in WSNs, highlighting the challenges posed by malware attacks due to the limited defense capabilities of sensor nodes, is discussed in [9,15]. In [10], the authors present a statistical methodology to predict malware propagation in WSNs using HJ-Biplot analysis combined with closed queuing networks. In [22], the authors investigate a new differential equation epidemic model (SUIQR) to describe malware propagation dynamics in WSNs. A mathematical model that describes malware propagation in WSNs, emphasizing the impact of different node types and network topology, is proposed in [18]. The work [21] investigates an improved SEIR model of fractional order to describe malware propagation in WSNs. In [19], the authors present the dynamics of a fractional-order malware propagation model in information networks, examining the stability of equilibrium points and the conditions for Hopf bifurcation. The authors in [14], discussed Securing WSN against multi-malware threats by fractional SE_1E_2VIQR model. However, this study did not consider the effect of varying the fractional order on the system. Furthermore, an analysis of the effect of the main parameters is missing.

3 Background Concepts

This section introduces all the essential elements required to comprehend the suggested fractional-order epidemic model comprehensively while emphasizing the foundational concepts from which this study's ideas emerge. These concepts serve as the basic building blocks upon which the entire proposal is established.

3.1 Fundamentals of Fractional Calculus

Initially, we present fundamental information on FDEs, including Caputo fractional derivatives and the existence and uniqueness of solutions to FDEs.

Definition 1. *[4, 8, 13] Suppose $y \in C^n([0,t])$. The Caputo fractional derivative of the function y, with an order $n-1 < \alpha < n$, where $n \in \mathbb{N}$, is given by:*

$$D_t^{\alpha} y(t) = \begin{cases} \frac{1}{\Gamma(n-\alpha)} \int_0^t \frac{y^{(n)}(z)}{(t-z)^{\alpha-n+1}} dz, & t > 0,\ 0 < n-1 < \alpha < n, \\ \frac{d^n y(t)}{dt^n}, & \alpha = n \in \mathbb{N}. \end{cases}$$

Lemma 1. *[3, 5] Take the following general FDE, which involves the Caputo derivative:*

$$\begin{aligned} &D_t^{\alpha} y(t) = g(t, y(t)), \quad 0 < \alpha \leq 1, \\ &y(t_0) = y_0, \end{aligned} \tag{1}$$

wherein $g(t, y(t)) : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ is a continuous function, called the vector field, related to y and dimension $m \geq 1$. If $g(t, y^*) = 0$, (1) has an equilibrium point at y^*.

Lemma 2. *[13] Suppose $g(t, y(t)) : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function related to y and dimension $m \geq 1$. If two positive constants Λ and δ exist such that $\|g(t, y(t))\| \leq \delta + \Lambda \|y(t)\|$ holds for every $(t, y) \in \mathbb{R} \times \mathbb{R}^m$, then a solution to the Caputo FDE (1) exists coming from the point (t_0, y_0).*

Remark 1. [13] If all conditions specified in Lemma 2 are satisfied and $\dfrac{\partial g(t, y(t))}{\partial y(t)}$ is continuous with respect to y, then the solution of the Caputo FDE (1) is certain to be unique.

Lemma 3. *[16] If $0 < \alpha \leq 1$, $y(t) \in C[a,b]$, and $D_t^{\alpha} y(t) \in C(a,b]$, then there exists $z \in [a,t]$ such that, according to the mean value theorem for integrals:*

$$y(t) = y(a) + \frac{1}{\alpha} (D_t^{\alpha} y)(z)(t-a)^{\alpha} \quad \text{for all } t \in (a,b].$$

Remark 2. Assume $0 < \alpha \leq 1$, $y(t) \in C[a,b]$, and $D_t^{\alpha} y(t) \in C(a,b]$. $y(t)$ is non-increasing if $D_t^{\alpha} y(t) \leq 0$ for all $t \in (a,b)$. On the other side, if $D_t^{\alpha} y(t) \geq 0$ for all $t \in (a,b)$, then $y(t) \in C[a,b]$ is non-decreasing for all $t \in [a,b]$.

3.2 Formulation of the Integer-Order Model

Recently, Awasthi et al. [1] designed a deterministic model for multi-malware propagation in WSNs considering five node classes: Susceptible, Exposed1, Exposed2, Infected, and Recovered. Such a model investigates two distinct forms of malware spread in WSNs, considering that the nodes fail or malfunction due to hardware or software issues when subjected to malware attacks. The authors suggested that states in a WSN can be classified into several categories at any time t. They have developed a series of ODEs to describe the dynamic process of malware spread in WSNs. The four distinct states constituting the proposed model are illustrated as follows:

- Susceptible S(t): It refers to the sensor nodes that remain unaffected by malware but are susceptible to infection at any moment.
- Exposed E(t): It denotes malware-infected sensor nodes that are not capable of spreading the malware despite their infection. However, over time, they distribute infection. Given the hypothesis that two distinct types of malware may simultaneously target the network, two categories of Exposed nodes, $E_1(t)$ and $E_2(t)$, are considered.
- infected I(t): It refers to the sensor nodes that are infected and have the potential to propagate the infection to adjacent sensor nodes.
- Recovered R(t): It signifies the category that includes sensor nodes within the network that have been restored following an Infected state.

The dynamical malware propagation formulated in the set of ODEs known as the SE_1E_2IR model is as follows [1]:

$$\begin{cases} \frac{dS}{dt} = b - \varphi\, S\, I - (\sigma + \omega)\, S \\ \frac{dE_1}{dt} = p\varphi\, S\, I - (\lambda_1 + \sigma)\, E_1 \\ \frac{dE_2}{dt} = q\varphi\, S\, I - (\lambda_2 + \sigma)\, E_2 \\ \frac{dI}{dt} = \lambda_1\, E_1 \, + \, \lambda_2\, E_2 \, - \, (\sigma + \gamma)\, I \\ \frac{dR}{dt} = \gamma\, I - \sigma\, R + \omega\, S \end{cases} \tag{2}$$

The system is characterized within the domain $\Gamma = \{(S, E_1, E_2, I, R) \in \mathbb{R}^5\}$ in which $S(0) = S_0$, $E_1(0) = 0$, $E_2(0) = 0$, $I(0) = I_0$, $R(0) = R_0$ for all $t \geq 0$. Because the model observes distinct states for sensor nodes, all state variables are positive for all $t \geq 0$. $N(t)$ uniformly distributed within the designated area L^2, denotes the total number of sensor nodes in the WSN at any given time t ($N(t) = S(t) + E_1(t) + E_2(t) + I(t) + R(t)$). The region covered by a sensor node of the WSN is πr^2, where r represents the communication radius. It should be mentioned that $p+q = 1$, $\lambda_1 > \lambda_2$, and $\varphi = \frac{\beta \pi r^2}{L^2}$. The most important parameters in the SE_1E_2IR model (2) are summarized in Table 1. Figure 1 shows the SE_1E_2IR state transition diagram.

The state of the model (2) at each time is not dependent on the previous history of the system; i.e., it is memoryless. These processes are known as Markovian since they do not have memory. As stated in the introduction, FDEs are invaluable tools for modeling the dynamics of WSNs. Models employing fractional-order derivatives suggest superior agreement between measured and simulated data compared with traditional models reliant on integer-order derivatives.

3.3 The Novel Fractional-Order SE_1E_2IR Model for Malware Propagation in WSNs

We aim to convert the model (2) into a fractional-order scenario that will integrate memory effects into the model. This study generalizes the prior SE_1E_2IR

Table 1. Parameters of the SE_1E_2IR model.

Parameters	Description
b	The addition rate of new nodes in WSN
σ	The failure rate of nodes caused by battery depletion or hardware/software problems
β	The coefficient of malware transmission in WSN
p	The rate of Susceptible nodes that became Exposed and enter into E_1 state
q	The rate of Susceptible nodes that became Exposed and enter into E_2 state
λ_1	The rate of Exposed state E_1 of nodes become Infected
λ_2	The rate of Exposed state E_2 of nodes become Infected
γ	The rate of recovery of Infected nodes
ω	The rate at which Susceptible nodes become immunized and move to the Recovered state R

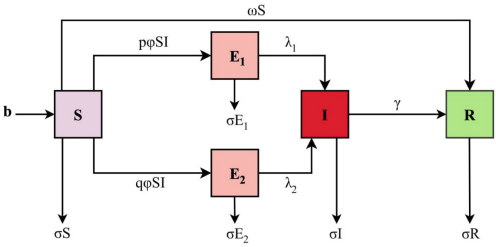


Fig. 1. State transition diagram of the model.

model, represented by ordinary differential equations, by incorporating the dynamics of states. In each case, we replace the conventional ordinary derivative with a fractional derivative to guarantee that both sides maintain equal dimensions. In particular, the following is a representation of the fractional-order malware propagation model that is based on the SE_1E_2IR model (2):

$$\begin{cases} D_t^\alpha S(t) = b^\alpha - \varphi^\alpha S\, I - (\sigma^\alpha + \omega^\alpha)\, S \\ D_t^\alpha E_1(t) = p^\alpha \varphi^\alpha S\, I - (\lambda_1^\alpha + \sigma^\alpha)\, E_1 \\ D_t^\alpha E_2(t) = q^\alpha \varphi^\alpha S\, I - (\lambda_2^\alpha + \sigma^\alpha)\, E_2 \\ D_t^\alpha I(t) = \lambda_1^\alpha E_1 + \lambda_2^\alpha E_2 - (\sigma^\alpha + \gamma^\alpha)\, I \\ D_t^\alpha R(t) = \gamma^\alpha I - \sigma^\alpha R + \omega^\alpha S \end{cases} \tag{3}$$

In (3), D_t^α is the Caputo fractional derivative utilized to integrate hereditary and historical characteristics into the new system. We have elevated all model parameters with dimension $\frac{1}{t^\alpha}$ to the exponent α to maintain dimensional consistency. The adjustment of the dimension mismatch is discussed in [12]. Therefore, both sides of the equations in (3) have identical dimensions. Henceforth, we will focus on the model (3).

4 Basic Results of the Model

To guarantee the validity of the model in monitoring populations of different classes, it is essential to ensure that all state variables remain nonnegative over the entire period of time. In this section, for the sake of convenience and simplicity, we shall refer to:

$$g(t,y(t)) = \begin{bmatrix} g_1(t,y(t)) \\ g_2(t,y(t)) \\ g_3(t,y(t)) \\ g_4(t,y(t)) \\ g_5(t,y(t)) \end{bmatrix} = \begin{bmatrix} b^\alpha - \varphi^\alpha\, S\, I - (\sigma^\alpha + \omega^\alpha)\, S \\ p^\alpha \varphi^\alpha\, S\, I - (\lambda_1^\alpha + \sigma^\alpha)\, E_1 \\ q^\alpha \varphi^\alpha\, S\, I - (\lambda_2^\alpha + \sigma^\alpha)\, E_2 \\ \lambda_1^\alpha\, E_1 + \lambda_2^\alpha\, E_2 - (\sigma^\alpha + \gamma^\alpha)\, I \\ \gamma^\alpha\, I - \sigma^\alpha\, R + \omega^\alpha\, S \end{bmatrix}. \tag{4}$$

4.1 Existence and Uniqueness

Theorem 1. *The fractional SE_1E_2IR model (3) has unique solution for all $t \geq 0$.*

Proof. We apply Lemma 2 to deduce that the fractional SE_1E_2IR model (3) has a solution for all $\mathbb{R}$. In the fractional SE_1E_2IR model (3), suppose $y(t) = (S(t), E_1(t), E_2(t), I(t), R(t))$. It is clear that $g(t, y(t))$ is continuous with respect to $y \in \mathbb{R}^5$. By using (4), we calculate

$\|g(t,y(t))\|_1 = \sum_{j=1}^{5} |g_j(t,y(t))|$ as follows:

$$\|g(t,y(t))\|_1 = |b^\alpha - \varphi^\alpha\, S\, I - (\sigma^\alpha + \omega^\alpha)\, S| + +|p^\alpha \varphi^\alpha\, S\, I - (\lambda_1^\alpha + \sigma^\alpha)\, E_1| + \\ +|q^\alpha \varphi^\alpha\, S\, I - (\lambda_2^\alpha + \sigma^\alpha)\, E_2| + +|\lambda_1^\alpha\, E_1 + \lambda_2^\alpha\, E_2 - (\sigma^\alpha + \gamma^\alpha)\, I| + +|\gamma^\alpha\, I - \sigma^\alpha\, R + \omega^\alpha\, S|. \tag{5}$$

By employing the triangular inequality, we obtain

$$\|g(t,y(t))\|_1 \leq b^\alpha + \varphi^\alpha(1 + p^\alpha + q^\alpha)|S||I| + +(\sigma^\alpha + 2\omega^\alpha)|S| + (2\lambda_1^\alpha + \sigma^\alpha)|E_1| + \\ +(2\lambda_2^\alpha + \sigma^\alpha)|E_2| + (\sigma^\alpha + 2\gamma^\alpha)\,|I| + \sigma^\alpha\,|R|. \tag{6}$$

Since, $N = S + E_1 + E_2 + I + R$ we have $S \leq N$. This implies

$$\|g(t,y(t))\|_1 \leq b^\alpha + (\varphi^\alpha(1 + p^\alpha + q^\alpha)N + +(\sigma^\alpha + 2\gamma^\alpha))|I| + (\sigma^\alpha + 2\omega^\alpha)|S| + \\ +(2\lambda_1^\alpha + \sigma^\alpha)|E_1| + (2\lambda_2^\alpha + \sigma^\alpha)|E_2| + \sigma^\alpha\,|R|. \tag{7}$$

Then, we add the subsequent positive terms to the right-hand side of inequality (7)

$$\begin{aligned} &(\varphi^\alpha(1 + p^\alpha + q^\alpha)N + 2(\gamma^\alpha + \lambda_1^\alpha + \lambda_2^\alpha))|S|, \\ &(\varphi^\alpha(1 + p^\alpha + q^\alpha)N + 2(\gamma^\alpha + \omega^\alpha + \lambda_2^\alpha))|E_1|, \\ &(\varphi^\alpha(1 + p^\alpha + q^\alpha)N + 2(\gamma^\alpha + \omega^\alpha + \lambda_1^\alpha))|E_2|, \\ &(2(\omega^\alpha + \lambda_1^\alpha + \lambda_2^\alpha))|I|, \\ &(\varphi^\alpha(1 + p^\alpha + q^\alpha)N + 2(\gamma^\alpha + \omega^\alpha + \lambda_1^\alpha + \lambda_2^\alpha))|R|. \end{aligned} \tag{8}$$

By adding the positive terms (8) to (7), we have:

$$\|g(t,y(t))\|_1 \leq b^\alpha + (\varphi^\alpha(1+p^\alpha+q^\alpha)N + 2(\gamma^\alpha+\omega^\alpha+\lambda_1^\alpha+\lambda_2^\alpha)+\sigma^\alpha)(|S|+|E_1|+|E_2| +|I|+|R|). \tag{9}$$

Since $\|y(t)\| = |S|+|E_1|+|E_2|+|I|+|R|$, we have

$$\|g(t,y(t))\|_1 \leq \delta + \Lambda\|y(t)\|, \quad \forall (t,y) \in \mathbb{R}\times\mathbb{R}^5 \tag{10}$$

where the positive constants are denoted as $b^\alpha = \delta$ and $\Lambda = (\varphi^\alpha(1+p^\alpha+q^\alpha)N+ 2(\gamma^\alpha+\omega^\alpha+\lambda_1^\alpha+\lambda_2^\alpha)+\sigma^\alpha)$. By utilizing Lemma 2, we conclude that the fractional SE_1E_2IR model (3) has a solution for all real numbers $\mathbb{R}$. Since $g(t,y(t)) \in \mathbb{R}^5$, $\dfrac{\partial g(t,y(t))}{\partial y(t)}$ is continuous with respect to y on $\mathbb{R}^5$. Utilizing Remark 1, we ascertain that the fractional SE_1E_2IR model (3) possesses a unique solution for all real numbers.

4.2 Positivity of the Solution

Theorem 2. *The permissible region* $\psi = \{(S,E_1,E_2,I,R) \in \mathbb{R}^5 : S \geq 0,\ E_1 \geq 0,\ E_2 \geq 0,\ I \geq 0,\ R \geq 0\}$ *for the fractional* SE_1E_2IR *model (3) is a positive invariant set.*

Proof. Suppose that $y(t) = (S(t),E_1(t),E_2(t),I(t),R(t))$ is the solution of the fractional SE_1E_2IR model (3). Now we compute $g(t,y(t))$ where $y(t) = (0,0,0,0,0)$. Since all the parameters in the fractional SE_1E_2IR model (3) are positive constants, for all $t \geq 0$, we have

$$g(t,y(t)) = \begin{bmatrix} D_t^\alpha S(t)|_{S=0} \\ D_t^\alpha E_1(t)|_{E_1=0} \\ D_t^\alpha E_2(t)|_{E_2=0} \\ D_t^\alpha I(t)|_{I=0} \\ D_t^\alpha R(t)|_{R=0} \end{bmatrix} = \begin{bmatrix} b^\alpha \\ p^\alpha\varphi^\alpha\, S\, I \\ q^\alpha\varphi^\alpha\, S\, I \\ \lambda_1^\alpha\, E_1 \,+\, \lambda_2^\alpha\, E_2 \\ \gamma^\alpha\, I + \omega^\alpha\, S \end{bmatrix} \geq 0. \tag{11}$$

Given that $g(t,y(t))$ is a continuous function, we deduce that $y(t) \in \psi$ and ψ is a positive invariant set of solution for the fractional SE_1E_2IR model (3) (According to Remark 2).

5 Main Results of the Model and Stability Analysis

This section will identify both equilibrium points. Furthermore, we will calculate the basic reproduction number and conduct a sensitivity analysis based on the free and endemic equilibrium points. This section concludes with an analysis of the stability of the endemic and free equilibrium points.

5.1 Free Equilibrium Points

The free equilibrium points of the system of the fractional SE_1E_2IR model (3) can be identified by solving the equations $D_t^\alpha S(t) = 0, D_t^\alpha E_1(t) = 0, D_t^\alpha E_2(t) = 0, D_t^\alpha I(t) = 0, D_t^\alpha R(t) = 0$. Consequently, we set

$$\begin{cases} b^\alpha - \varphi^\alpha S\, I - (\sigma^\alpha + \omega^\alpha)\, S = 0, \\ p^\alpha \varphi^\alpha S\, I - (\lambda_1^\alpha + \sigma^\alpha)\, E_1 = 0, \\ q^\alpha \varphi^\alpha S\, I - (\lambda_2^\alpha + \sigma^\alpha)\, E_2 = 0, \\ \lambda_1^\alpha E_1 + \lambda_2^\alpha E_2 - (\sigma^\alpha + \gamma^\alpha)\, I = 0, \\ \gamma^\alpha I - \sigma^\alpha R + \omega^\alpha S = 0. \end{cases} \tag{12}$$

A straightforward computation will determine the free equilibrium points (Ξ_0) as follows:

$$\Xi_0 = \left(\frac{b^\alpha}{(\sigma^\alpha + \omega^\alpha)}, 0, 0, 0, \frac{\omega^\alpha b^\alpha}{\sigma^\alpha(\sigma^\alpha + \omega^\alpha)}\right). \tag{13}$$

5.2 Basic Reproduction Number

The basic reproduction number, R_0, is the primary parameter for categorizing the spread of malware in WSN. R_0 represents the total number of secondary Infected nodes caused by a single infected node over a certain period. Hence, R_0 is a dimensionless metric identifying the probability of malware propagation. Various methodologies exist for determining R_0 in the context of epidemic spread [6]. This study employs the next-generation matrix technique. Just E_1, E_2, and I are presently directly engaged in the spread of malware. Upon examination of the issue, we have delineated three classes, denoted as (E_1, E_2, I), to determine the parameter R_0. Consequently, we recall the fractional SE_1E_2IR model (3) here:

$$\begin{cases} D_t^\alpha E_1(t) = p^\alpha \varphi^\alpha S\, I - (\lambda_1^\alpha + \sigma^\alpha)\, E_1 \\ D_t^\alpha E_2(t) = q^\alpha \varphi^\alpha S\, I - (\lambda_2^\alpha + \sigma^\alpha)\, E_2 \\ D_t^\alpha I(t) = \lambda_1^\alpha E_1 + \lambda_2^\alpha E_2 - (\sigma^\alpha + \gamma^\alpha)\, I \end{cases} \tag{14}$$

Employing the approach (the next generation matrix [6]), the system (14) can be expressed as $D_t^\alpha X = P(X) - Q(X)$, where

$$X = \begin{bmatrix} E_1 \\ E_2 \\ I \end{bmatrix}, \quad P(X) = \begin{bmatrix} p^\alpha \varphi^\alpha S\, I \\ q^\alpha \varphi^\alpha S\, I \\ \lambda_1^\alpha E_1 + \lambda_2^\alpha E_2 \end{bmatrix}, Q(X) = \begin{bmatrix} (\lambda_1^\alpha + \sigma^\alpha)\, E_1 \\ (\lambda_2^\alpha + \sigma^\alpha)\, E_2 \\ (\sigma^\alpha + \gamma^\alpha)\, I \end{bmatrix}.$$

The Jacobian matrices of $P(X)$ and $Q(X)$ are calculated as follows:

$$\mathcal{J}(P(X)) = \begin{bmatrix} 0 & 0 & p^\alpha \varphi^\alpha S \\ 0 & 0 & q^\alpha \varphi^\alpha S \\ \lambda_1^\alpha & \lambda_2^\alpha & 0 \end{bmatrix},$$

and

$$\mathcal{J}(Q(X)) = \begin{bmatrix} (\lambda_1^\alpha + \sigma^\alpha) & 0 & 0 \\ 0 & (\lambda_2^\alpha + \sigma^\alpha) & 0 \\ 0 & 0 & (\sigma^\alpha + \gamma^\alpha) \end{bmatrix}.$$

The Jacobian matrices of $P(X)$ and $Q(X)$ at the free equilibrium points (Ξ_0) are calculated according to (13), as follows:

$$K := \mathcal{J}(P|_{\Xi_0}) = \begin{bmatrix} 0 & 0 & \frac{p^\alpha \varphi^\alpha b^\alpha}{(\sigma^\alpha + \omega^\alpha)} \\ 0 & 0 & \frac{q^\alpha \varphi^\alpha b^\alpha}{(\sigma^\alpha + \omega^\alpha)} \\ \lambda_1^\alpha & \lambda_2^\alpha & 0 \end{bmatrix}, H := \mathcal{J}(Q|_{\Xi_0}) = \begin{bmatrix} (\lambda_1^\alpha + \sigma^\alpha) & 0 & 0 \\ 0 & (\lambda_2^\alpha + \sigma^\alpha) & 0 \\ 0 & 0 & (\sigma^\alpha + \gamma^\alpha) \end{bmatrix}. \tag{15}$$

The inverse of H is given by

$$H^{-1} = \begin{bmatrix} \frac{1}{(\lambda_1^\alpha + \sigma^\alpha)} & 0 & 0 \\ 0 & \frac{1}{(\lambda_2^\alpha + \sigma^\alpha)} & 0 \\ 0 & 0 & \frac{1}{(\sigma^\alpha + \gamma^\alpha)} \end{bmatrix}. \tag{16}$$

The next generation matrix $M = KH^{-1}$ is:

$$M = \begin{bmatrix} 0 & 0 & \frac{p^\alpha \varphi^\alpha b^\alpha}{(\sigma^\alpha + \omega^\alpha)(\sigma^\alpha + \gamma^\alpha)} \\ 0 & 0 & \frac{q^\alpha \varphi^\alpha b^\alpha}{(\sigma^\alpha + \omega^\alpha)(\sigma^\alpha + \gamma^\alpha)} \\ \frac{\lambda_1^\alpha}{(\lambda_1^\alpha + \sigma^\alpha)} & \frac{\lambda_2^\alpha}{(\lambda_2^\alpha + \sigma^\alpha)} & 0 \end{bmatrix}. \tag{17}$$

Consequently, the basic reproduction number R_0 (which indicates the largest eigenvalue) is calculated as follows:

$$R_0 = \frac{b^\alpha \varphi^\alpha}{(\sigma^\alpha + \omega^\alpha)(\sigma^\alpha + \gamma^\alpha)} \Big(\frac{p^\alpha \lambda_1^\alpha}{(\lambda_1^\alpha + \sigma^\alpha)} + \frac{q^\alpha \lambda_2^\alpha}{(\lambda_2^\alpha + \sigma^\alpha)} \Big). \tag{18}$$

The scalar R_0 is a dimensionless threshold value commonly referred to as the basic reproduction number when $\alpha = 1$.

5.3 Endemic Equilibrium Points

If $R_0 > 1$, there exists another kind of equilibrium, termed endemic equilibrium $\Xi^* = (S^*, E_1^*, E_2^*, I^*, R^*)$, that occurs when the infection remains continuously present throughout the system. The endemic equilibrium points of the system of the fractional SE_1E_2IR model (3) can be identified by solving equations $D_t^\alpha S^*(t) = 0, D_t^\alpha E_1^*(t) = 0, D_t^\alpha E_2^*(t) = 0, D_t^\alpha I^*(t) = 0, D_t^\alpha R^*(t) = 0$. A straightforward computation will determine the endemic equilibrium points Ξ^* as follows:

$$\begin{aligned} \Xi^* = \Big(& \frac{b^\alpha}{(\sigma^\alpha + \omega^\alpha) R_0}, \frac{(R_0 - 1) p^\alpha b^\alpha}{R_0 (\sigma^\alpha + \lambda_1^\alpha)}, \frac{(R_0 - 1) q^\alpha b^\alpha}{R_0 (\sigma^\alpha + \lambda_2^\alpha)}, \Big(\frac{(R_0 - 1)(\sigma^\alpha + \omega^\alpha)}{\varphi^\alpha} \Big), \\ & \frac{1}{\sigma^\alpha} \Big(\frac{\omega^\alpha b^\alpha}{R_0 (\sigma^\alpha + \omega^\alpha)} + \frac{\gamma^\alpha (\sigma^\alpha + \omega^\alpha)(R_0 - 1)}{\varphi^\alpha} \Big) \Big) \end{aligned} \tag{19}$$

5.4 Stability Analysis of the Fractional SE_1E_2IR Model in Free Equilibrium Points

Regarding the introduced fractional SE_1E_2IR model (3), we should investigate the stability of equilibrium points. This section seeks to identify the local and global stability of free equilibrium points. Indeed, this subsection evaluates local and global stability concerning the free equilibrium points Ξ_0.

Definition 2. *A steady state solution y^* of an FDE system is called locally asymptotically stable if each solution $y(t)$ that initiates close to y^* converges to y^* as $t \longrightarrow \infty$.*

Definition 3. *A steady state solution y^* of an FDE system is globally asymptotically stable if every solution of FDE converges to y^* as $t \longrightarrow \infty$.*

Theorem 3. *The fractional SE_1E_2IR model (3) demonstrates local asymptotic stability at Ξ_0 when $R_0 < 1$ and unstable when $R_0 > 1$ (.i.e., if all eigenvalues of the matrix (5.4) are negative).*

Proof. The Jacobian matrix of the fractional SE_1E_2IR model (3) at the free equilibrium Ξ_0, can be expressed as follows:

$$\mathcal{J}|_{\Xi_0} = \begin{bmatrix} -(\sigma^\alpha + \omega^\alpha) & 0 & 0 & -\varphi^\alpha S_0 & 0 \\ 0 & -(\lambda_1^\alpha + \sigma^\alpha) & 0 & p^\alpha \varphi^\alpha S_0 & 0 \\ 0 & 0 & -(\lambda_2^\alpha + \sigma^\alpha) & q^\alpha \varphi^\alpha S_0 & 0 \\ 0 & \lambda_1^\alpha & \lambda_2^\alpha & -(\sigma^\alpha + \gamma^\alpha) & 0 \\ \omega^\alpha & 0 & 0 & \gamma^\alpha & -\sigma^\alpha \end{bmatrix}$$

The characteristic equation of the matrix $\mathcal{J}$ is:

$$P_j = (\eta + (\sigma^\alpha + \omega^\alpha))\,(\eta + \sigma^\alpha)\,(\eta^3 + \xi_1 \eta^2 + \xi_2 \eta + \xi_3)$$

where

$$\begin{cases} \xi_1 = Tr(A) = (\lambda_1^\alpha + \lambda_2^\alpha) + 3\sigma^\alpha + \gamma^\alpha, \\ \xi_2 = (A_{ji}A_{ij} - A_{ii}A_{jj}) = (\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha) + (\lambda_1^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha) + (\lambda_1^\alpha + \sigma^\alpha)(\lambda_2^\alpha + \sigma^\alpha) - \\ (\lambda_1^\alpha p^\alpha \varphi^\alpha S_0 - \lambda_2^\alpha q^\alpha \varphi^\alpha S_0), \\ \xi_3 = det(A) = p^\alpha \varphi^\alpha S_0 \lambda_1^\alpha (\lambda_2^\alpha + \sigma^\alpha) + q^\alpha \varphi^\alpha S_0 \lambda_2^\alpha (\lambda_1^\alpha + \sigma^\alpha) - (\lambda_1^\alpha + \sigma^\alpha)(\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha). \end{cases}$$

where A is the matrix as follows:

$$A = \begin{bmatrix} -(\lambda_1^\alpha + \sigma^\alpha) & 0 & p^\alpha \varphi^\alpha S_0 \\ 0 & -(\lambda_2^\alpha + \sigma^\alpha) & q^\alpha \varphi^\alpha S_0 \\ \lambda_1^\alpha & \lambda_2^\alpha & -(\sigma^\alpha + \gamma^\alpha) \end{bmatrix}$$

After precise verification, it is obvious that for $R_0 < 1$, all coefficients ξ_1, ξ_2, and ξ_3 of the cubic equation $\eta^3 + \xi_1 \eta^2 + \xi_2 \eta + \xi_3$ are strictly positive. Consequently, the characteristic equation has five roots, $\eta_1, \eta_2, \eta_3, \eta_4$, and η_5, all of which have negative real parts. This fact, in turn, indicates that all five eigenvalues of the matrix $\mathcal{J}|_{\Xi_0}$ possess negative real parts; hence, the fractional SE_1E_2IR model (3) demonstrates local stability at the malware-free equilibrium points Ξ_0. Given these findings, the proof is considered complete.

Theorem 4. *If $R_0 \leq 1$, then the free equilibrium points Ξ_0 are globally asymptotically stable.*

Proof. According to the Lyapunov stability theorem [11], if the system possesses a definite positive function, referred to as $\mathcal{L}(t)$, and its full derivative $D_t^\alpha \mathcal{L}(t) \leq 0$ along Ξ_0 with respect to t, consequently, the system is globally asymptotically stable at Ξ_0. We can select some parameters from the model and its state equations to construct a Lyapunov function. Suppose

$$\mathcal{L} = \Delta\Big(\lambda_1^\alpha(\lambda_2^\alpha + \sigma^\alpha)E_1 + \lambda_2^\alpha(\lambda_1^\alpha + \sigma^\alpha)E_2 + (\lambda_1^\alpha + \sigma^\alpha)(\lambda_2^\alpha + \sigma^\alpha)I\Big)$$

be a positive Lyapunov function, where

$$\Delta = \frac{b}{(\sigma^\alpha + \lambda_1^\alpha)(\sigma^\alpha + \lambda_2^\alpha)(\sigma^\alpha + \gamma^\alpha)}.$$

The fractional derivative of $\mathcal{L}$ with respect to t is

$$D_t^\alpha \mathcal{L}(t) = \Delta\Big(\lambda_1^\alpha(\lambda_2^\alpha + \sigma^\alpha)\, D_t^\alpha E_1(t) + \lambda_2^\alpha(\lambda_1^\alpha + \sigma^\alpha)\, D_t^\alpha E_2(t) + (\lambda_1^\alpha + \sigma^\alpha)(\lambda_2^\alpha + \sigma^\alpha)\, D_t^\alpha I(t)\Big) =$$
$$\Delta\Big(\lambda_1^\alpha(\lambda_2^\alpha + \sigma^\alpha)\,(p^\alpha\varphi^\alpha\, S\, I - (\lambda_1^\alpha + \sigma^\alpha)\, E_1) + \lambda_2^\alpha(\lambda_1^\alpha + \sigma^\alpha)\,(q^\alpha\varphi^\alpha\, S\, I - (\lambda_2^\alpha + \sigma^\alpha)\, E_2) + (\lambda_1^\alpha + \sigma^\alpha)$$
$$(\lambda_2^\alpha + \sigma^\alpha)\,(\lambda_1^\alpha\, E_1 + \lambda_2^\alpha\, E_2 - (\sigma^\alpha + \gamma^\alpha)\, I)\Big)$$

Utilizing the basic reproduction number and $S = \frac{b^\alpha}{(\sigma^\alpha + \omega^\alpha)}$, we obtain:

$$D_t^\alpha \mathcal{L}(t) = (R_0 - 1)I$$

After analyzing the function $D_t^\alpha \mathcal{L}(t)$, we deduce that $D_t^\alpha \mathcal{L}(t) \leq 0$ along Ξ_0 if and only if $R_0 \leq 1$. If the function $I(t) = 0$, then the maximum value of the operator $D_t^\alpha \mathcal{L}(t)$ is also zero. Consequently, the global stability at Ξ_0 adheres to the LaSalle invariance principle [17] if $R_0 \leq 1$.

5.5 Stability Analysis of the Fractional SE_1E_2IR Model in Endemic Equilibrium Points

This part evaluates the proposed system stability concerning the endemic equilibrium points Ξ^*.

Theorem 5. *The fractional SE_1E_2IR model (3) demonstrates local asymptotic stability at Ξ^* if all eigenvalues of the matrix (20) are negative.*

Proof. The Jacobian matrix of the fractional SE_1E_2IR model (3), for malware propagation in WSNs, at the malware-endemic equilibrium points Ξ^*, can be expressed as follows:

$$\mathcal{J}|_{\Xi^*} = \begin{bmatrix} -(\varphi^\alpha I^* + \sigma^\alpha + \omega^\alpha) & 0 & 0 & -\varphi^\alpha S^* & 0 \\ p^\alpha\varphi^\alpha I^* & -(\lambda_1^\alpha + \sigma^\alpha) & 0 & p^\alpha\varphi^\alpha S^* & 0 \\ q^\alpha\varphi^\alpha I^* & 0 & -(\lambda_2^\alpha + \sigma^\alpha) & q^\alpha\varphi^\alpha S^* & 0 \\ 0 & \lambda_1^\alpha & \lambda_2^\alpha & -(\sigma^\alpha + \gamma^\alpha) & 0 \\ \omega^\alpha & 0 & 0 & \gamma^\alpha & -\sigma^\alpha \end{bmatrix} \tag{20}$$

The characteristic equation of the matrix $\mathcal{J}|_{\Xi^*}$ is:

$$P_j^* = (\eta + \sigma^\alpha)\left(\eta^4 + \xi_1^*\eta^3 + \xi_2^*\eta^2 + (\xi_3^*\eta) + \xi_4^*\right)$$

where

$$\begin{cases} \xi_1^* = Tr(A^*) = \lambda_1^\alpha + \lambda_2^\alpha + 4\sigma^\alpha + \gamma^\alpha + \omega^\alpha + \varphi^\alpha I^*, \\ \xi_2^* = (\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha) + (2\sigma^\alpha + \lambda_1^\alpha + \omega^\alpha + \varphi^\alpha I^*)(2\sigma^\alpha + \lambda_2^\alpha + \gamma^\alpha)(\lambda_1^\alpha + \sigma^\alpha)(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*) - \\ p^\alpha\varphi^\alpha S^*\lambda_1^\alpha - q^\alpha\varphi^\alpha S^*\lambda_2^\alpha \\ \xi_3^* = (\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha)(2\sigma^\alpha + \lambda_1^\alpha + \omega^\alpha + \varphi^\alpha I^*) + (2\sigma^\alpha + \lambda_2^\alpha + \gamma^\alpha)(\lambda_1^\alpha + \sigma^\alpha)(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*) + p^\alpha(\varphi^\alpha)^2 \\ S^*I^*\lambda_1^\alpha + q^\alpha(\varphi^\alpha)^2S^*I^*\lambda_2^\alpha - p^\alpha\varphi^\alpha S^*\lambda_1^\alpha(2\sigma^\alpha + \lambda_2^\alpha + \omega^\alpha + \varphi^\alpha I^*) - q^\alpha\varphi^\alpha S^*\lambda_2^\alpha(2\sigma^\alpha + \lambda_1^\alpha + \omega^\alpha + \varphi^\alpha I^*) \\ \xi_4^* = det(A^*) = (\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*)(\lambda_1^\alpha + \sigma^\alpha)(\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha) + p^\alpha(\varphi^\alpha)^2S^*I^*\lambda_1^\alpha(\lambda_2^\alpha + \sigma^\alpha) + q^\alpha(\varphi^\alpha)^2S^*I^*\lambda_2^\alpha \\ (\lambda_1^\alpha + \sigma^\alpha) - p^\alpha\varphi^\alpha S^*\lambda_1^\alpha(\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*) - q^\alpha\varphi^\alpha S^*\lambda_2^\alpha(\lambda_1^\alpha + \sigma^\alpha)(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*) \end{cases}$$

where A^* is the matrix as follows:

$$A^* = \begin{bmatrix} -(\varphi^\alpha I^* + \sigma^\alpha + \omega^\alpha) & 0 & 0 & -\varphi^\alpha S^* \\ p^\alpha\varphi^\alpha I^* & -(\lambda_1^\alpha + \sigma^\alpha) & 0 & p^\alpha\varphi^\alpha S^* \\ q^\alpha\varphi^\alpha I^* & 0 & -(\lambda_2^\alpha + \sigma^\alpha) & q^\alpha\varphi^\alpha S^* \\ 0 & \lambda_1^\alpha & \lambda_2^\alpha & -(\sigma^\alpha + \gamma^\alpha) \end{bmatrix}.$$

All coefficients $\xi_1^*, \xi_2^*, \xi_3^*$, and ξ_4^* of equation $\eta^4 + \xi_1^*\eta^3 + \xi_2^*\eta^2 + (\xi_3^*\eta) + \xi_4^*$ should be positive for satisfying Routh-Hurwitz criteria. It is obvious that $\xi_1^* > 0$, for other coefficients ξ_2^*, ξ_3^*, and ξ_4^*, we have:
$\xi_2^* > 0$ if:

$$(\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha) + (2\sigma^\alpha + \lambda_1^\alpha + \omega^\alpha + \varphi^\alpha I^*)(2\sigma^\alpha + \lambda_2^\alpha + \gamma^\alpha)(\lambda_1^\alpha + \sigma^\alpha)(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*) > p^\alpha\varphi^\alpha S^*\lambda_1^\alpha + q^\alpha\varphi^\alpha S^*\lambda_2^\alpha,$$

$\xi_3^* > 0$ if:

$$(\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha)(2\sigma^\alpha + \lambda_1^\alpha + \omega^\alpha + \varphi^\alpha I^*) + (2\sigma^\alpha + \lambda_2^\alpha + \gamma^\alpha)(\lambda_1^\alpha + \sigma^\alpha)(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*) + p^\alpha(\varphi^\alpha)^2 S^*I^*\lambda_1^\alpha + q^\alpha(\varphi^\alpha)^2S^*I^*\lambda_2^\alpha > p^\alpha\varphi^\alpha S^*\lambda_1^\alpha(2\sigma^\alpha + \lambda_2^\alpha + \omega^\alpha + \varphi^\alpha I^*) + q^\alpha\varphi^\alpha S^*\lambda_2^\alpha(2\sigma^\alpha + \lambda_1^\alpha + \omega^\alpha + \varphi^\alpha I^*),$$

$\xi_4^* > 0$ if:

$$(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*)(\lambda_1^\alpha + \sigma^\alpha)(\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \gamma^\alpha) + p^\alpha(\varphi^\alpha)^2S^*I^*\lambda_1^\alpha(\lambda_2^\alpha + \sigma^\alpha) + q^\alpha(\varphi^\alpha)^2S^*I^*\lambda_2^\alpha(\lambda_1^\alpha + \sigma^\alpha) > p^\alpha\varphi^\alpha S^*\lambda_1^\alpha(\lambda_2^\alpha + \sigma^\alpha)(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*) + q^\alpha\varphi^\alpha S^*\lambda_2^\alpha(\lambda_1^\alpha + \sigma^\alpha)(\sigma^\alpha + \omega^\alpha + \varphi^\alpha I^*).$$

Consequently, The characteristic equation has five roots, $\eta_1, \eta_2, \eta_3, \eta_4$, and η_5, all of which have negative real parts. Therefore, this fact indicates that all five eigenvalues of the matrix $\mathcal{J}|_{\Xi^*}$ include negative real parts, thereby confirming the local stability of the fractional SE_1E_2IR model (3) at the malware-endemic equilibrium points Ξ^*. Given these observations, the proof is considered complete.

Theorem 6. *If $R_0 \geq 1$, then the endemic equilibrium points Ξ^* are globally asymptotically stable.*

Proof. When $R_0 > 1$, suppose $\Psi(x) = x - x^* - x^* ln\frac{x}{x^*}$ and construct the Liapunov function as

$$\mathcal{L}^*(t) = \tau_1\Psi(S(t)) + \tau_2\Psi(E_1(t)) + \tau_3\Psi(E_2(t)) + \tau_4\Psi(I(t)) + \tau_5\Psi(R(t))$$

where $\tau_i > 0, i = 1, 2, ..., 5$. The fractional derivative of $\mathcal{L}^*$ with respect to t is

$$D^\alpha \mathcal{L}^*(t) = D^\alpha \Big(\tau_1 \Psi(S(t)) + \tau_2 \Psi(E_1(t)) + \tau_3 \Psi(E_2(t)) + \tau_4 \Psi(I(t)) + \tau_5 \Psi(R(t))\Big)$$

Utilizing (3), we have

$$D_t^\alpha \mathcal{L}^*(t) \leq \tfrac{S-S^*}{S} D^\alpha S(t) + \tfrac{E_1-E_1^*}{E_1} D^\alpha E_1(t) + \tfrac{E_2-E_2^*}{E_2} D^\alpha E_2(t) + \tfrac{I-I^*}{I} D^\alpha I(t) + \tfrac{R-R^*}{R} D^\alpha R(t)$$

After analyzing the function $D_t^\alpha \mathcal{L}^*(t)$, we deduce that $D_t^\alpha \mathcal{L}^*(t) \leq 0$ along Ξ^* if and only if $R_0 \geq 1$. If the function $S(t) = S^*, E_1(t) = E_1^*, E_2(t) = E_2^*, I(t) = I^*, R(t) = R^*$, then the maximum value of the operator $D_t^\alpha \mathcal{L}^*(t)$ is also zero. Consequently, the global stability at Ξ^* adheres to the LaSalle invariance principle [17] if $R_0 \geq 1$. Therefore, Ξ^* is globally attractive. Combined with the local stability of Ξ^*, it can be seen that Ξ^* is globally asymptotically stable.

6 Simulations and Analysis of the Malware Propagation Model in WSNs

In this section, based on its capabilities, we will use the FABM method to investigate stability and confirm the obtained theoretical results.

Case 1. Assume that $b = 0.4$, $\beta = 0.001$, $\lambda_1 = 0.002$, $\lambda_2 = 0.001$, $\sigma = 0.001$, $\gamma = 0.003$, $\omega = 0.004$, $r = 1$, $L = 12$, and $p = 0.7$. The values of $R_0 = 0.2691$ for $\alpha = 1$ and $R_0 = 0.1019$ for $\alpha = 0.80$ can be computed using (18). When $R_0 < 1$, the various sensor nodes are displayed in Fig. 2. The number of Susceptible nodes decreases as the number of nodes in other categories rises in WSN, according to the analysis of Fig. 2 for case 1. However, when $R_0 < 1$, the number of I, E_1, E_2, and R nodes gradually starts to decline and gets closer to zero. The Infected nodes in the WSN eventually vanish. As a result, the malware is eliminated from the WSN automatically. Based on Fig. 2, all solutions converge to Ξ_0 as $t \to \infty$ for case 1. Therefore, according to the graphical method, this fulfills Theorems 3 and 4 and shows that for $R_0 < 1$, the WSN will be both locally and globally stable.

Case 2. Let us to consider $b = 0.4$, $\beta = 0.002$, $\lambda_1 = 0.002$, $\lambda_2 = 0.001$, $\sigma = 0.0001$, $\gamma = 0.003$, $\omega = 0.004$, $r = 1.5$, $L = 10$, and $p = 0.7$. The values of $R_0 = 4.1795$ for $\alpha = 1$ and $R_0 = 1.4695$ for $\alpha = 0.80$ can both be calculated using (18). The various sensor nodes are shown in Fig. 3 when $R_0 > 1$. The malware will consistently be present in WSN for case 2, as seen in Fig. 3. The solutions $R(t), I(t), E_1(t)$, and $E_2(t)$ for the case 2 are clearly visible in Fig. 3. These solutions move away from Ξ_0 as $t \to \infty$. Consequently, the graphical method complies with Theorem 3, which states that the WSN will become unstable when $R_0 > 1$. Under these circumstances, malware will propagate within the WSN;

therefore, malware continuously persists in the system. Theorems 5 and 6 are validated by this result. It is observed that the number of Infected nodes rises quickly in the early stages, then begins to decrease before stabilizing. Malware is still spreading among network nodes, demonstrating how infection and recovery rates affect malware propagation in WSNs. This is also consistent with the main conclusion drawn from Theorems 5 and 6, which states that the system is finally at the endemic equilibrium point Ξ^*.

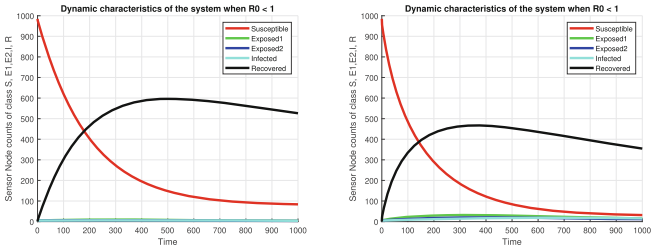


Fig. 2. Status Variables when $R_0 < 1$ (Left: $\alpha = 1$, and Right: $\alpha = 0.80$).

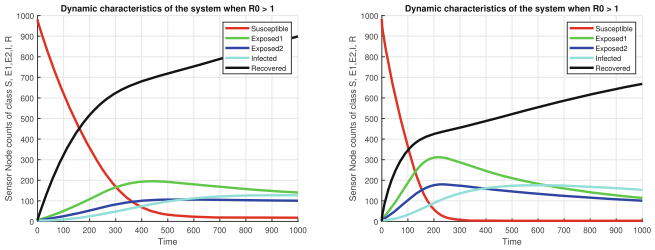


Fig. 3. Status variables when $R_0 > 1$ (Left: $\alpha = 1$, and Right: $\alpha = 0.80$).

6.1 Analyzing the Effect of Fractional Derivative α

Using the following parameter values, we examine how the fractional derivative order α affects the system when $R_0 < 1$. Let us to consider $b = 0.4$, $\beta = 0.001$, $\lambda_1 = 0.002$, $\lambda_2 = 0.001$, $\sigma = 0.001$, $\gamma = 0.003$, $\omega = 0.004$, $r = 1$, $L = 12$, and $p = 0.7$ (Case 1). The dynamic relation between time and S, E_1, E_2, I, and R nodes is depicted in Fig. 4. Figure 4 makes it clear that, for a range of α values, the status variables connected to different nodes in WSN decrease over time. Furthermore, the image makes it clear that, in contrast to a lesser value of α, the malware leaves the WSN very quickly as α converges to 1. Observing the

relation between time and Recovered nodes shows that recovery speed reduces as α gets closer to 1. As illustrated in Fig. 4, all solutions converge to Ξ_0 as $t \to \infty$ when α converges to 1. The WSN will, therefore, show both local and global asymptotic stability for a range of values of α when $R_0 < 1$, according to the graphic approach, which concludes that the Theorems 3 and 4 are satisfied. Additionally, Fig. 4 clarifies that the FABM method yields positive answers and converges toward its accurate steady state.

We also investigate how the system is affected when $R_0 > 1$ by the fractional derivative order α. The parameter values used in this investigation are $b = 0.4$, $\beta = 0.002$, $\lambda_1 = 0.002$, $\lambda_2 = 0.001$, $\sigma = 0.0001$, $\gamma = 0.003$, $\omega = 0.004$, $r = 1.5$, $L = 10$, and $p = 0.7$ (Case 2). The dynamic relationship between time and S, E_1, E_2, I, and R nodes is shown in Fig. 5. For different values of α, Fig. 5 shows that the number of S nodes in WSN declines with time. However, given varying values of α, the number of E_1, E_2, I, and R nodes increases over time. Moreover, Fig. 5 clarifies that the malware stays on the system indefinitely as α gets closer to 1. The malware will spread throughout the WSN in this situation. The number of Recovered nodes rises as α gets closer to 1, according to the correlation between time and the number of Recovered nodes. As seen in Fig. 5, all solutions that do not converge to Ξ_0 as $t \to \infty$ do so when α goes to 1. As $t \to \infty$ for varying values of α, the four solutions $R(t), I(t)$, $E_1(t)$, and $E_2(t)$ move away from Ξ_0. The graphic approach thus leads to the conclusion that Theorem 3 is satisfied, meaning that when $R_0 > 1$, the WSN will become unstable for different values of α. In this situation, malware spreads in the WSN, where it continuously persists. This result validates Theorems 5 and 6. In this case, the FABM technique also converges toward its accurate steady state and provides positive solutions, as seen in Fig. 5.

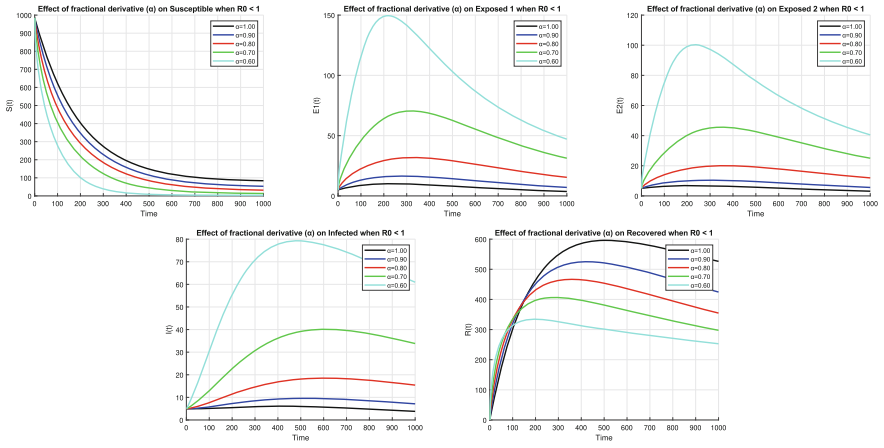


Fig. 4. Analyzing the effect of α when $R_0 < 1$.

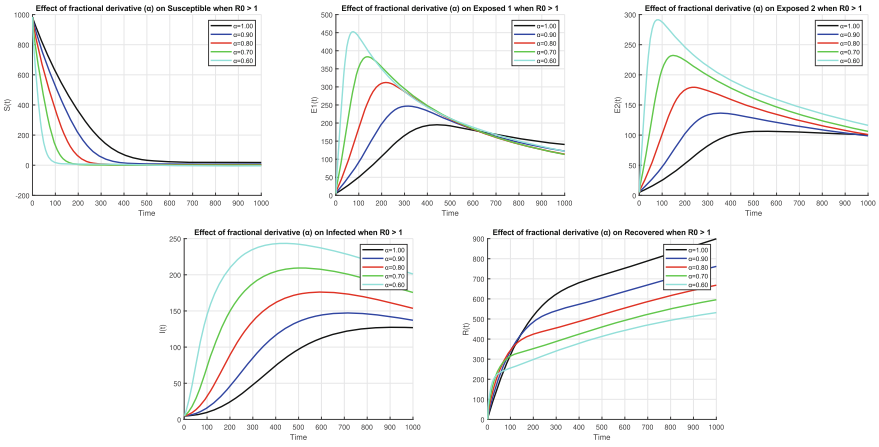


Fig. 5. Analyzing the effect of α when $R_0 > 1$.

7 Conclusion

This work illustrated the applicability of fractional derivatives in investigating the propagation dynamics of multi-malware infections in WSNs using epidemiological modeling. The equilibrium points for a malware-free state have been identified, and R_0 has been calculated to assess the system dynamics and analyze the stability requirements of the system. The analysis has demonstrated that when $R_0 < 1$, the WSN will be stabilized in a malware-free equilibrium; however, if $R_0 > 1$, malware will continue to exist in the WSN. The network's stability has been analyzed under various situations, with the corresponding theorems and their proofs provided. The FABM approach was employed to confirm the theoretical findings and identify the problem of malware attacks. Simulation results indicated that the proposed fractional-order propagation model outperforms the integer-order propagation one.

Acknowledgements. This work was partially supported by SERICS (PE00000014) under PNRR MUR program funded by EU-NGEU.

References

1. Awasthi, S., et al.: An epidemic model for the investigation of multi-malware attack in wireless sensor network. IET Commun. **17**(11), 1274–1287 (2023). https://doi.org/10.1049/cmu2.12622
2. Brauer, F., Castillo-Chavez, C., Feng, Z.: Mathematical Models in Epidemiology. TAM, vol. 69. Springer, New York (2019). https://doi.org/10.1007/978-1-4939-9828-9

3. Conte, D., Farsimadan, E., Moradi, L., Palmieri, F., Paternoster, B.: Time-delay fractional optimal control problems: a survey based on methodology. In: Abdel Wahab, M. (ed.) Proceedings of the 8th International Conference on Fracture, Fatigue and Wear, pp. 325–337. Springer Singapore, Singapore (2021). https://doi.org/10.1007/978-981-15-9893-7_23
4. Conte, D., Farsimadan, E., Moradi, L., Palmieri, F., Paternoster, B.: A galerkin approach for fractional delay differential equations using hybrid chelyshkov basis functions. In: Gervasi, O. (ed.) Computational Science and Its Applications – ICCSA 2022, pp. 130–142. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-10522-7_10
5. Conte, D., Farsimadan, E., Moradi, L., Palmieri, F., Paternoster, B.: Numerical solution of delay volterra functional integral equations with variable bounds. Appl. Numer. Math. **200**, 148–161 (2024). https://doi.org/10.1016/j.apnum.2023.05.002
6. De Santis, A., Farsimadan, E., Moradi, L., Palmieri, F.: Generalized defensive modeling of fake news propagation in social networks using fractional differential equations. IEEE Trans. Comput. Soc. Syst. **12**(2), 622–634 (2025). https://doi.org/10.1109/TCSS.2024.3492097
7. Dong, N.P., Giang, N.L., Long, H.V.: Interconnected takagi-sugeno system and fractional sirs malware propagation model for stabilization of wireless sensor networks. Inf. Sci. **670**, 120620 (2024). https://doi.org/10.1016/j.ins.2024.120620
8. Farsimadan, E., Moradi, L., Conte, D., Paternoster, B., Palmieri, F.: Comparison between protein-protein interaction networks $cd4^+t$ and $cd8^+t$ and a numerical approach for fractional HIV infection of $cd4^+t$ cells. In: Gervasi, O. (ed.) Computational Science and Its Applications – ICCSA 2021, pp. 78–94. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86653-2_6
9. Farsimadan, E., Moradi, L., Palmieri, F.: A general study on the malware propagation models in wireless sensor networks. In: Gervasi, O. (ed.) Computational Science and Its Applications – ICCSA 2024 Workshops, pp. 83–99. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-65223-3_6
10. Frutos-Bernal, E., Rodríguez-Rosa, M., Anciones-Polo, M., Martín-del Rey, A.: Analyzing malware propagation on wireless sensor networks: a new approach using queueing theory and HJ-biplot with a sirs model. Mathematics **12**(1) (2024). https://doi.org/10.3390/math12010135
11. Geng, L., Zheng, H., Qiao, G., Geng, L., Wang, K.: Online public opinion dissemination model and simulation under media intervention from different perspectives. Chaos, Solitons Fractals **166**, 112959 (2023). https://doi.org/10.1016/j.chaos.2022.112959
12. Graef, J.R., Kong, L., Ledoan, A., Wang, M.: Stability analysis of a fractional online social network model. Math. Comput. Simul. **178**, 625–645 (2020). https://doi.org/10.1016/j.matcom.2020.07.012
13. Lin, W.: Global existence theory and chaos control of fractional differential equations. J. Math. Anal. Appl. **332**(1), 709–726 (2007). https://doi.org/10.1016/j.jmaa.2006.10.040
14. Mishra, J., Srivastava, P.K., Srivastava, V.: Securing wireless sensor network against multi-malware threats: a comprehensive modeling and analysis approach. SN Comput. Sci. **6**(3), 236 (2025)
15. Moradi, L., Farsimadan, E., D'Angelo, G., Carpentieri, B., Palmieri, F.: Epidemic analysis of the propagation of multiple malware infectious in wireless sensor networks. In: Barolli, L. (ed.) Advanced Information Networking and Applications, pp. 140–148. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-87784-1_13

16. Odibat, Z.M., Shawagfeh, N.T.: Generalized Taylor's formula. Appl. Math. Comput. **186**(1), 286–293 (2007). https://doi.org/10.1016/j.amc.2006.07.102
17. LaSalle, J.P.: The Stability of Dynamical Systems. SIAM (1976). https://doi.org/10.1137/1.9781611970432
18. Martín-del Rey, A.: A novel model for malware propagation on wireless sensor networks. Math. Biosci. Eng. **21**(3), 3967–3998 (2024). https://doi.org/10.3934/mbe.2024176
19. Shi, X., Luo, A., Chen, X., Huang, Y., Huang, C., Yin, X.: The dynamical behaviors of a fractional-order malware propagation model in information networks. Mathematics **12**(23) (2024). https://doi.org/10.3390/math12233814
20. Verma, C., Gupta, C.P.: Effect of vaccination on stability of wireless sensor network against malware attack: an epidemiological model. SN Comput. Sci. **5**(2), 240 (2024)
21. Zhou, Y., Liu, B.T., Zhou, K., Shen, S.F.: Malware propagation model of fractional order, optimal control strategy and simulations. Front. Phys. **11** (2023). https://doi.org/10.3389/fphy.2023.1201053
22. Zhou, Y., Wang, Y., Zhou, K., Shen, S.F., Ma, W.X.: Dynamical behaviors of an epidemic model for malware propagation in wireless sensor networks. Front. Phys. **11** (2023). https://doi.org/10.3389/fphy.2023.1198410

Sensitivity Analysis of the Multi Malware Attacks Model in Wireless Sensor Networks

Leila Moradi[1(✉)], Eslam Farsimadan[1], Marcello Trovati[2], and Francesco Palmieri[1]

[1] Department of Computer Science, University of Salerno, 84084 Fisciano, Italy
{lmoradi,efarsimadan,fpalmieri}@unisa.it
[2] School of Business, University of Lancashire, Preston, Lancashire, UK
mtrovati@lancashire.ac.uk

Abstract. Wireless Sensor Networks (WSNs) technology is extensively utilized in various applications necessitating monitoring and control functionalities. Nonetheless, some unresolved difficulties impede its effective deployment, primarily related to security concerns. This study investigates and analyzes the parameter sensitivity of the Susceptible-Exposed1-Exposed2-Infected-Recovered (SE_1E_2IR) model, derived from the classical Susceptible-Exposed-Infected-Recovered (SEIR) epidemic theory and used for the early detection of multi-malware activities in WSNs. In this model, two exposed states have been considered based on the assumption of two different types of malware attacks in the network. The primary SE_1E_2IR model develops a mechanism for the earlier detection of malware attack incidents in WSNs. The model is structured upon a system of ordinary differential equations. The basic reproduction number is also determined as a crucial factor for characterizing malware propagation within the network. It can be helpful to identify the circumstances in which the network remains almost safe or the risk of a malware outbreak. This work offers a fundamental comprehension of sensitivity analysis for the epidemic SE_1E_2IR model. Furthermore, it examines the efficacy of the parameters and analyzes the impact of different model parameters on malware spread in WSNs. Then, an analytical comparison of the suggested model with existing models is presented. Numerical simulations are run in MATLAB to augment the theoretical analyses.

Keywords: Differential Equations · Wireless sensor networks · Malware propagation · Basic reproductive number · Sensitivity analysis

1 Introduction

Wireless Sensor Networks (WSNs) are gaining an ever-increasing interest in modern communication scenarios due to their essential roles in cyber-physical systems and their related applications. They are now extensively used in environmental monitoring, intelligent transportation, fire prevention, rescue management, security monitoring, traffic control, target tracking, and many other

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 406–422, 2027.
https://doi.org/10.1007/978-3-032-30521-3_25

domains due to their simple configuration, extreme adaptability, and inherently scalable nature [2,14,31]. However, several known issues and open challenges, mainly due to hardware constraints and limitations, can adversely affect their secure deployment on mission-critical infrastructures. Indeed, WSNs are still susceptible to various menaces, Sybil [16,35], wormhole [9] and selective forwarding [37] attacks, as well as malware propagation [1,25] and related malicious activities. Among these menaces, malware propagation can be considered particularly dangerous, mostly in large-scale WSNs, since any infected sensor node can easily spread malware to its neighbors, leading to the rapid compromise of the entire infrastructure. To easily harm the entire network, an attacker may target particular nodes (e.g., cluster heads or gateways), inject malware components, and rely on the native propagation mechanism characterizing these infrastructures (i.e., clustering properties associated with scale-free dynamics) to immediately infect the whole system. In addition, radio communications, pervasively used in modern WSNs, offer extremely suitable circumstances for massive malware propagation, mainly when ad-hoc mesh connection architectures are involved. New-generation malware propagates autonomously through the mesh, rapidly affecting the whole network without human involvement. Accordingly, knowledge of malware propagation dynamics is crucial to preserving network security and integrity. Usually, more sophisticated WSNs can deploy protection strategies against the spread of malware, together with detection and recovery mechanisms that can discover and repair malware-infected nodes. However, managing such a mix of proactive and reactive strategies requires non-negligible activities on communication channels, resulting in additional energy consumption and, hence, reduced lifetimes of sensor nodes. It is straightforward to consider that a better understanding of the malware diffusion process in WSNs can be extremely helpful in driving more targeted protection strategies that are able to effectively reduce the overhead by minimizing the overall power loss [38]. Some recent research efforts on the propagation of malware in WSNs are presented in [11,12,18].

Since a variety of epidemic models for infectious diseases have been proposed, the dynamics of malware propagation in networks may also be explained by these models. Therefore, in order to describe the process of malware propagation in network systems, researchers developed theoretical models that took inspiration from existing ideas of biological viral transmission across populations. Nevertheless, limited effort has been devoted to malware propagation models that are specific to WSNs. In several recent studies, researchers have used epidemic models to study how malware spreads in WSNs [4,23,39]. Since WSNs have special dynamic properties, the models of spreading epidemics in the human population cannot be simply applied to investigate malware propagation in WSNs. Furthermore, the behaviors of different malware affecting WSNs varied significantly and unexpectedly. In particular, common models only consider a single type of malware infection active at any time, and more realistic schemes considering the possibility of multiple simultaneous malware attacks (multi-malware models) can be extremely interesting for more realistic modeling. On the other hand, the majority of models that have been suggested to study the spread of mal-

ware in these networks rely on Ordinary Differential Equations (ODEs), which are an ineffective tool for describing the behaviors of a complicated networked system [5–7,10]. Researchers have suggested multiple epidemiological models to examine the stability and effectiveness of WSN.

Sensitivity analysis of model parameters is crucial in mathematical model formulation, as it clarifies the relationship between alterations in model outputs and fluctuations in model inputs [8]. This study aims to serve as an initial guide for individuals interested in exploring parameter sensitivity in the mathematical modeling of malware dissemination in WSNs through ODEs. The direct differential method is a fast way to do sensitivity analysis. It analyzes local sensitivities, allowing researchers to gain a thorough understanding of the impacts of slight parameter variations on model outcomes. In the subsequent sections of this study, we analyze the local and global sensitivity of the SE_1E_2IR model.

The rest of this paper is organized as follows: Sect. 2 reports some related work. In Sects. 3, description and formulation of the proposed model are presented. Section 4 discusses the basic properties of the model. The sensitivity analysis of the model parameters is covered in Sect. 5. The simulation outcomes of the modified model and the discussion of the achieved results are presented in Sect. 6. Finally, Sect. 7 presents the conclusion and the upcoming work.

2 Related Work

Recent studies on malware propagation on networked systems, such as WSNs, are based on epidemic theories, which, in turn, rely on mathematical models where differential equations play an essential role. A mathematical model for an epidemic phenomenon is fundamentally built on a few significant assumptions regarding the problem of interest. For instance, concerning malware propagation, the hypotheses are entirely correlated with the characteristics associated with malware infection and network connections, allowing its diffusion among involved entities. By employing these mathematical models, investigations on the dynamics of malware propagation in WSNs have yielded notable advancements. Most of the available propagation models, including SIS, SIRS, SEIR, SEIRS, SITR, SNIRD, HSIRD, SEIRV, and VSCIQRS are established based on the conventional SIR model.

Ojha et al. [22] introduced a Susceptible-Exposed-Infected-Recovered-Susceptible (SEIRS) model to address the limitations of previous models. The suggested SEIRS model contains a finite communication radius and the corresponding node density for WSNs, offering a method to mitigate the propagation of worms. Upadhyay and Kumari [29] presented an energy-efficient e-epidemic Susceptible-Infected-Terminally Infected-Recovered (SITR) model to examine the attacking behaviors of worms in WSNs utilizing a cytoid-type functional response. Geetha et al. [13] suggested a modified SEIRS model using background noise as an additional variable. The effectiveness of the model has been validated through Lyapunov-based analysis. Shen et al. [26,27] proposed a heterogeneous Susceptible-iNsidious-Infected-Recovered-Dysfunctional(SNIRD) model

that takes into account the connectivity of multiple heterogeneous sensor nodes while also reflecting the properties of malware concealment and dysfunctional nodes. Moreover, they suggested another Heterogeneous Susceptible-Infected-Removed-Dead (HSIRD) model for heterogeneous WSNs, in which they determined the epidemic threshold and conducted stability assessments for equilibrium points.

Regarding agent-based modeling, Nwokoye and Umeh [20] introduced the agent equivalent of the mathematical Susceptible-Exposed-Infected-Recovered-Susceptible with a Vaccination class (SEIRS-V) model. Furthermore, Xu et al. [33] established the Susceptible-Exposed-Infected-Recovered with a Failed type (SEIRS-F) agent-based model to deal with malware propagation and component reliability in WSNs.

Han et al. [15] suggested an approach to intrusion detection in WSNs utilizing an autoregressive model and game theory to minimize energy consumption in case of deploying Intrusion Detection Systems. Besides decreasing energy usage, this model can forecast attack targets. Singh et al. [28] presented an enhanced Susceptible-Exposed-Infected-Recovered-Vaccination (SEIRV) model to address the effects of the radius of communication and density of nodes on malware spreading in WSNs, aiming at identifying the essential conditions for the continuous spread of malware within the network. Wu et al. [32] conducted experiments on malware propagation in heterogeneous WSNs leveraging the Markov chain technique and infectious disease theories, revealing that network availability correlates with the vulnerability, distribution, and degree of network nodes. Acarali et al. [1] developed a Susceptible-Infected-Susceptible (SIS) malware propagation model to examine the dynamics of the spread of botnet within Internet of Things (IoT)-based WSNs, disclosing that node distribution density and energy constraints influence the botnet generation process.

Wu et al. [30] studied a virus propagation model to analyze the propagation of a virus in the Social Internet of Things (SIoT), taking into account the variable number of associated devices. They employed game theory to mitigate the virus propagation among the devices. Zhang et al. [36] suggested a method based on epidemic theory to mitigate the propagation of malware in WSNs. It involves sensor network heterogeneity by considering sensor nodes and sink nodes. They examined the impact of malware attacks on several kinds of topologies and proposed a framework for setting up, distributing, and preserving WSNs. Ye et al. [34] developed a Susceptible-Infected-Basically Recovered-Completely Recovered (SIR_1R_2) theory that examined the influence of secondary immunity mechanisms on malware propagation in WSNs. Zhou et al. [38] leveraged game theory to examine the malware spread dynamics in WSNs. Furthermore, the Nash equilibrium technique is utilized to assess the propagation patterns of malware in such networks. The correlation between the steady-state infection ratio and game characteristics and the presence of malware in WSNs is investigated. Cellular automata have also been used to simulate malware spread in WSNs and validate the accuracy of the proposed theory with simulation outcomes. Liu et al. [17] proposed an alternative SEIR model to characterize the spreading

of mutated viruses in WSNs. The basic reproductive number is calculated by implementing the next-generation matrix approach. The stability analysis of the system has also been conducted, resulting in the identification of two equilibrium points: global and local. They tried to propose a practical control approach for managing the spread of viruses in WSNs.

Although many studies investigate the behavior of malware propagation in WSNs, few have investigated the sensitivity analysis of multi-malware propagation in WSNs. Motivated by this shortage, we consider epidemic model for multi-malware attacks in WSNs based on ODEs and we discuss the local and global sensitivity analysis methods.

3 Background Concepts

Recently, Awasthi et al. [3] designed a deterministic model for multi-malware propagation in WSNs considering five node classes: Susceptible, Exposed 1, Exposed 2, Infected, and Recovered. Such a model investigates two distinct forms of malware spread in WSNs, considering that the nodes fail or malfunction due to hardware or software issues when subjected to malware attacks. At first, it was considered that all nodes were susceptible to malware attacks. The operational efficiency of sensor nodes is decreased as they become infected. This model includes two latent phases, revealing that malware requires a certain time to become really infectious. The average duration spent in the E_i subclass is determined as $\frac{1}{(\lambda_i+\sigma)}$ [24]. The authors suggested that states in a WSN can be classified into several categories at any time t. They have developed a series of ODEs to describe the dynamic process of malware spread in WSNs. The four distinct states constituting the proposed model are illustrated as follows:

- Susceptible $S(t)$: It refers to the sensor nodes that remain unaffected by malware but are susceptible to infection at any moment.
- Exposed $E(t)$: It denotes malware-infected sensor nodes that are not capable of spreading the malware despite their infection. However, over time, they distribute infection. Given the hypothesis that two distinct types of malware may simultaneously target the network, two categories of exposed nodes, $E_1(t)$ and $E_2(t)$, are considered. Given the hypothesis that two distinct types of malware may simultaneously target the network, two categories of exposed nodes, $E_1(t)$ and $E_2(t)$, are considered.
- Infected $I(t)$: It refers to the sensor nodes that are infected and have the potential to propagate the infection to adjacent sensor nodes.
- Recovered $R(t)$: It signifies the category that includes sensor nodes within the network that have been restored following an infected state.

The dynamical malware propagation formulated in the set of ODEs known as the SE_1E_2IR model is as follows [3]:

$$\begin{cases} \frac{dS}{dt} = b - \varphi\, S\, I - (\sigma + \omega)\, S \\ \frac{dE_1}{dt} = p\varphi\, S\, I - (\lambda_1 + \sigma)\, E_1 \\ \frac{dE_2}{dt} = q\varphi\, S\, I - (\lambda_2 + \sigma)\, E_2 \\ \frac{dI}{dt} = \lambda_1\, E_1 \;+\; \lambda_2\, E_2 \;-\; (\sigma + \gamma)\, I \\ \frac{dR}{dt} = \gamma\, I - \sigma\, R + \omega\, S \end{cases} \tag{1}$$

The system is characterized within the domain $\Gamma = \{(S, E_1, E_2, I, R) \in \mathbb{R}^5\}$ in which $S(0) = S_0$, $E_1(0) = 0$, $E_2(0) = 0$, $I(0) = I_0$, $R(0) = R_0$ for all $t \geq 0$. Because the model observes distinct states for sensor nodes, all state variables are positive for all $t \geq 0$.

$N(t)$ uniformly distributed within the designated area L^2, denotes the total number of sensor nodes in the WSN at any given time t ($N(t) = S(t) + E_1(t) + E_2(t) + I(t) + R(t)$). The region covered by a sensor node of the WSN is πr^2, where r represents the communication radius. It should be mentioned that $p + q = 1$, $\lambda_1 > \lambda_2$, and $\varphi = \frac{\beta \pi r^2}{L^2}$. The most important parameters in the SE_1E_2IR model (1) are summarized in Table 1.

Table 1. Parameters of the SE_1E_2IR model.

Parameters	Description
b	The addition rate of new nodes in WSN
σ	The failure rate of nodes caused by battery depletion or hardware/software problems
β	The coefficient of malware transmission in WSN
p	The rate of susceptible nodes that became exposed and enter into E_1 state
q	The rate of susceptible nodes that became exposed and enter into E_2 state
λ_1	The rate of exposed state E_1 of nodes become infectious
λ_2	The rate of exposed state E_2 of nodes become infectious
γ	The rate of recovery of infectious nodes
ω	The rate at which susceptible nodes become immunized and move to the recovered state R

4 Main Results of the Model and Stability Analysis

The analysis of sensitivity investigates the impact of parameters on these equilibria, consequently validating models through the comparison of actual and predicted values. We will calculate the basic reproduction number and conduct a sensitivity analysis based on the free equilibrium points.

4.1 Free Equilibrium Points

The free equilibrium points of the system of the SE_1E_2IR model (1) can be identified by solving the equations $\frac{dS}{dt} = 0, \frac{dE_1}{dt} = 0, \frac{dE_2}{dt} = 0, \frac{dI}{dt} = 0, \frac{dR}{dt} = 0$. Consequently, we set

$$\begin{cases} b - \varphi\, S\, I - (\sigma + \omega)\, S = 0, \\ p\varphi\, S\, I - (\lambda_1 + \sigma)\, E_1 = 0, \\ q\varphi\, S\, I - (\lambda_2 + \sigma)\, E_2 = 0, \\ \lambda_1\, E_1 \, + \, \lambda_2\, E_2 \, - \, (\sigma + \gamma)\, I = 0, \\ \gamma\, I - \sigma\, R + \omega\, S = 0. \end{cases} \tag{2}$$

A straightforward computation will determine the free equilibrium points (Ξ_0) as follows:

$$\Xi_0 = (\frac{b}{(\sigma + \omega)}, 0\,, 0,\, 0,\, \frac{\omega b}{\sigma(\sigma + \omega)}). \tag{3}$$

4.2 Basic Reproduction Number

The basic reproduction number, R_0, is the primary parameter for categorizing the spread of malware in WSN. R_0 represents the total number of secondary infected nodes caused by a single infected individual over a certain period. Hence, R_0 is a dimensionless metric identifying the probability of malware propagation. Various methodologies exist for determining R_0 in the context of epidemic spread. This study employs the next-generation matrix technique. Just E_1, E_2, and I are presently directly engaged in the spread of malware. Upon examination of the issue, we have delineated three classes, denoted as (E_1, E_2, I), to determine the parameter R_0. Consequently, we recall the SE_1E_2IR model (1) here:

$$\begin{cases} \dfrac{dE_1}{dt} = p\varphi\, S\, I - (\lambda_1 + \sigma)\, E_1 \\ \dfrac{dE_2}{dt} = q\varphi\, S\, I - (\lambda_2 + \sigma)\, E_2 \\ \dfrac{dI}{dt} = \lambda_1\, E_1 \, + \, \lambda_2\, E_2 \, - \, (\sigma + \gamma)\, I \end{cases} \tag{4}$$

Employing the approach (the next generation matrix [8]), the system (4) can be expressed as $\dfrac{dX}{dt} = P(X) - Q(X)$, where

$$X = \begin{bmatrix} E_1 \\ E_2 \\ I \end{bmatrix}, \quad P(X) = \begin{bmatrix} p\varphi\, S\, I \\ q\varphi\, S\, I \\ \lambda_1\, E_1 \, + \, \lambda_2\, E_2 \end{bmatrix}, Q(X) = \begin{bmatrix} (\lambda_1 + \sigma)\, E_1 \\ (\lambda_2 + \sigma)\, E_2 \\ (\sigma + \gamma)\, I \end{bmatrix}.$$

The Jacobian matrices of $P(X)$ and $Q(X)$ are calculated as follows:

$$\mathcal{J}(P(X)) = \begin{bmatrix} 0 & 0 & p\varphi S \\ 0 & 0 & q\varphi S \\ \lambda_1 & \lambda_2 & 0 \end{bmatrix}, \mathcal{J}(Q(X)) = \begin{bmatrix} (\lambda_1 + \sigma) & 0 & 0 \\ 0 & (\lambda_2 + \sigma) & 0 \\ 0 & 0 & (\sigma + \gamma) \end{bmatrix}.$$

The Jacobian matrices of $P(X)$ and $Q(X)$ at the free equilibrium points (Ξ_0) are calculated according to (3), as follows:

$$K := \mathcal{J}(P|_{\Xi_0}) = \begin{bmatrix} 0 & 0 & \frac{p\varphi b}{(\sigma+\omega)} \\ 0 & 0 & \frac{q\varphi b}{(\sigma+\omega)} \\ \lambda_1 & \lambda_2 & 0 \end{bmatrix}, H := \mathcal{J}(Q|_{\Xi_0}) = \begin{bmatrix} (\lambda_1+\sigma) & 0 & 0 \\ 0 & (\lambda_2+\sigma) & 0 \\ 0 & 0 & (\sigma+\gamma) \end{bmatrix}.$$

The inverse of H is given by

$$H^{-1} = \begin{bmatrix} \frac{1}{(\lambda_1+\sigma)} & 0 & 0 \\ 0 & \frac{1}{(\lambda_2+\sigma)} & 0 \\ 0 & 0 & \frac{1}{(\sigma+\gamma)} \end{bmatrix}. \tag{5}$$

The next generation matrix $M = KH^{-1}$ is:

$$M = \begin{bmatrix} 0 & 0 & \frac{p\varphi b}{(\sigma+\omega)(\sigma+\gamma)} \\ 0 & 0 & \frac{q\varphi b}{(\sigma+\omega)(\sigma+\gamma)} \\ \frac{\lambda_1}{(\lambda_1+\sigma)} & \frac{\lambda_2}{(\lambda_2+\sigma)} & 0 \end{bmatrix}. \tag{6}$$

Consequently, the basic reproduction number R_0 (which indicates the largest eigenvalue) is calculated as follows:

$$R_0 = \frac{b\varphi}{(\sigma+\omega)(\sigma+\gamma)}\Big(\frac{p\lambda_1}{(\lambda_1+\sigma)} + \frac{q\lambda_2}{(\lambda_2+\sigma)}\Big). \tag{7}$$

5 Efficiency of the Parameters

This part investigates the efficacy of the parameters and explores how various model parameters affect malware propagation in WSNs. This implies, in sensitivity analysis, a minor alteration of any parameter can result in a significant variation in the malware's ability to propagate over WSNs. Here, we consider two invaluable tools for efficiency analysis: local and global sensitivity. It is possible to identify the most influential parameters of the proposed model using both local and global sensitivity analysis.

Local Sensitivity Analysis. Analyzing the sensitivity of R_0 is crucial for effectively suppressing malware propagation in WSNs, as R_0 is an essential factor in its transmission. There is an index that represents the sensitivity of R_0 to changes in the parameter Υ. Thus, by employing partial derivatives, we can calculate the variations in all parameters in the R_0 equation.

The formula demonstrating the R_0 sensitivity index concerning the parameter Υ is:

$$\Gamma_\Upsilon = (\frac{\partial R_0}{\partial \Upsilon})(\frac{\Upsilon}{R_0}) \tag{8}$$

where Υ is a parameter that describes R_0. A negative (positive) index signifies that an augmentation in the parameter Υ results in a reduction (an increase) of R_0. The sensitivity indices for the parameters $b, \beta, \lambda_1, \lambda_2, \sigma, \omega$, and γ can be derived from the basic reproduction number, as shown in Table 2. It is clear that among these six indices, R_0 exhibits the most significant sensitivity to variations in ω, γ, and σ.

Table 2. Sensitivity indices.

Parameters	The sensitivity index of R_0 with respect to the parameters	Indexes of sensitivity
b	1	Positive
β	1	Positive
λ_1	$\frac{bp\varphi\sigma\lambda_1}{R_0(\sigma+\gamma)(\sigma+\omega)(\sigma+\lambda_1)^2}$	Positive
λ_2	$\frac{bq\varphi\sigma\lambda_2}{R_0(\sigma+\gamma)(\sigma+\omega)(\sigma+\lambda_2)^2}$	Positive
σ	$-\sigma\left[\frac{p\frac{\lambda_1}{(\lambda_1+\sigma)^2}+q\frac{\lambda_2}{(\lambda_2+\sigma)^2}}{p\frac{\lambda_1}{\lambda_1+\sigma}+q\frac{\lambda_2}{\lambda_2+\sigma}}+\frac{2\sigma+\gamma+\omega}{(\sigma+\gamma)(\sigma+\omega)}\right]$	Negative
ω	$\frac{-\omega}{(\sigma+\omega)}$	Negative
γ	$\frac{-\gamma}{(\gamma+\sigma)}$	Negative

Global Sensitivity Analysis. The most crucial model parameters can be determined using global sensitivity analysis. Indeed, the most important characteristics are identified by combining Latin Hypercube Sampling (LHS) with the Partial Rank Correlation Coefficient (PRCC) [19]. The LHS is a stratified Monte Carlo sampling technique that makes equal probability intervals of sample size N out of the parameter distribution. Each interval is sampled just once, without any replacement. Additionally, every parameter's whole range is investigated. After ranking the inputs and outputs, PRCC indices are calculated. The PRCC analysis identifies the model's most influential parameters, some of which have a positive correlation and others a negative one with the number of infected nodes. This indicates that as the value of any parameter with a positive correlation increases, the number of infected nodes increases. Conversely, the number of infected nodes drops if the value of a parameter with negative correlation increases.

6 Simulations and Analysis of the Malware Propagation Model in WSNs

The simulation outcomes of the modified model and the discussion of the achieved results are presented in this section. In this section, based on its capabilities, we will use the ABM method to investigate stability and confirm the

obtained theoretical results. This method enables the precise computation of numerical solutions over time.

6.1 Local and Global Sensitivity Analysis of the Model Parameters

ω, σ, and γ are the most sensitive to changes among the parameters that characterize R_0. Let us describe how the parameters ω, σ, and γ of the model could control the spread of malware in WSN. The parameters $b = 0.4$, $\beta = 0.002$, $\lambda_1 = 0.002$, $\lambda_2 = 0.001$, $\sigma = 0.0001$, $\gamma = 0.003$, $\omega = 0.004$, $r = 1.5$, $L = 10$, and $p = 0.7$ will be fixed in order to examine the impact of the parameters ω, σ, and γ. Only the sensitive parameters will be changed.

First, we fix the parameter value $\omega = 0.004, 0.006, 0.008, 0.01$. A higher recovery rate of susceptible nodes ω leads to removing infected nodes from the WSN, as shown in Fig. 1. This figure illustrates how malware will become less prevalent throughout the network as recovery rates increase. This strategy can effectively stop malware from spreading. As illustrated in Fig. 1, the peak emerges earlier with fewer infected nodes in the WSN as the parameter ω value increases. The changes can be seen in Fig. 1, where the dynamic relation between time and infected nodes and time and recovered nodes are shown separately in Fig. 1. The dynamic relation between infected and recovered nodes is also shown in this figure for a range of ω values.

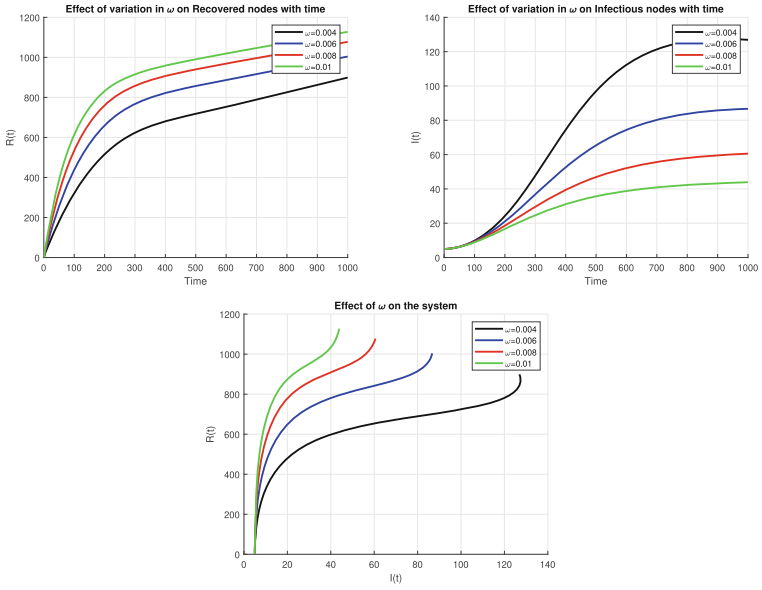


Fig. 1. Effect of ω on the system.

We change the parameter $\sigma = 0.0001, 0.0003, 0.0006, 0.0009$ to investigate the impact of σ on WSN. There are fewer infected and recovered nodes in WSN

as σ increases, which reduces the probability of becoming susceptible to malware spread. In WSN, the peak happens earlier when there are fewer infected nodes. The dynamic relation between the recovered, infected nodes with time is shown separately in Fig. 2. Under the given conditions, the number of recovered and infected nodes within WSN increases. It is evident from the data in Fig. 2 that the number of infected and recovered nodes in WSN is negatively correlated with the death rate. As a result, there will be a much lower chance of malware spreading throughout the network in this case. However, this removal procedure does not prevent infected nodes from spreading malware throughout the network. They may be categorized as spreaders because they are susceptible to being victimized by infected nodes. In particular, the dynamic link between time and infected people is shown in Fig. 2. In contrast, the other ones show the dynamic relation between infected-recovered nodes for varying values of σ, respectively. By changing the recovery rate γ to $\gamma = 0.003, 0.005, 0.007$, malware occurrence is significantly and quickly decreased. Increasing the parameter γ affects the number of infected nodes, slows the speed of the actively infected nodes, and may even prevent malware from spreading. An increase in γ similarly raises the number of recovered nodes over time from the WSN, as shown in Fig. 3. This method can stop malware from spreading within WSN. The number of recovered nodes increases as the parameter γ as a recovery rate increases in less time, as seen in the right image of Fig. 3. To be more specific, the dynamic relation between time and recovered nodes is depicted in Fig. 3. In contrast, the dynamic relation between infected and recovered nodes is also depicted in this figure for a range of γ values.

Figure 4 depicts the PRCC analysis findings of the most crucial parameters of the model, $b, \lambda_1, \lambda_2, \sigma, \omega, \gamma$, and β. Some of these parameters have a positive correlation with the number of infected nodes, while others have a negative correlation. According to the findings of the PRCC analysis, the proportion of nodes that propagate malware is positively correlated with $b, \lambda_1, \lambda_2, \sigma$, and β; when these parameters increase, the proportion of nodes that are affected by malware will also rise. However, σ has an almost insignificant effect in comparison with the other parameters. Moreover, it can be seen that b and β have a greater effect on the proportion of nodes that spread malware than λ_1 and λ_2. On the other side, ω, γ is negatively correlated with the proportion of nodes that spread malware. γ has an insignificant effect on it, whereas ω has a strong negative correlation. The aforementioned experiment outcomes are compatible with the results of the local sensitivity parameters shown in Figs. 1–3, demonstrating that the current proposed model is applicable to multi-malware propagation in WSNs.

6.2 Comparative Analysis with Existing Model

This part presents a comparison analysis between the proposed model and two existing models [3,21]. In the following case studies, we examine two distinct scenarios.

Case 1: For comparative study, in this case, we consider and fix all the parameters as: $b = 0.4, \beta = 0.002, \lambda_1 = 0.002, \lambda_2 = 0.001, \sigma = 0.0001, \gamma =$

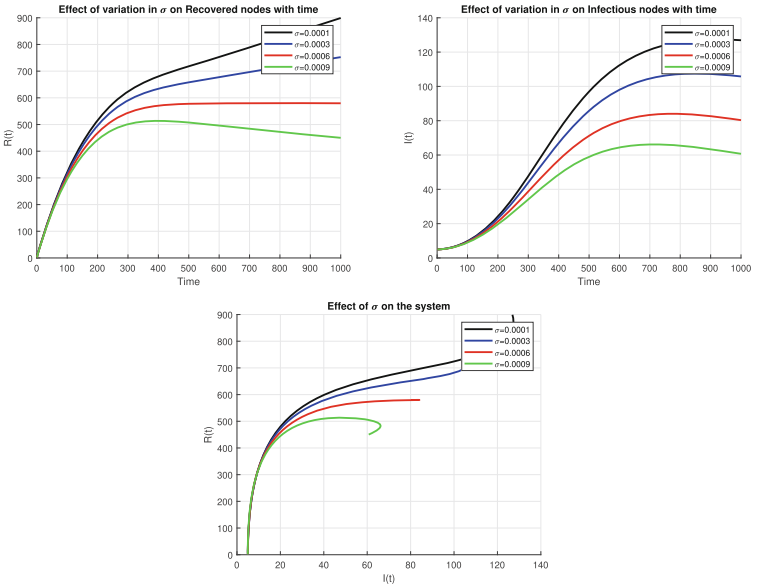


Fig. 2. Effect of σ on the system.

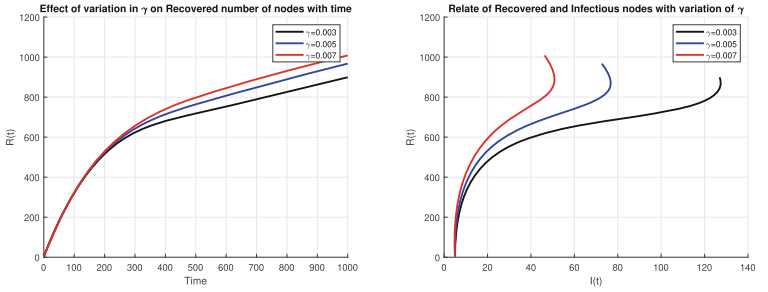


Fig. 3. Effect of recovery rate γ on the system.

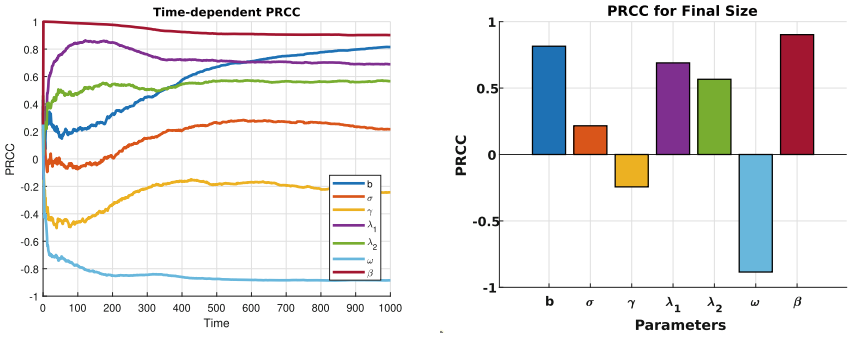


Fig. 4. PRCC of the parameters in the dynamical model.

$0.003, \omega = 0.004, L = 10, p = 0.7, N = 1000$, and we alter just the value of communication radius as $r = 1.0$ and $r = 1.5$. Table 3 summarizes the comparison of existing and proposed models. Considering Table 3, compared to the existing models, in the proposed model, a smaller number of susceptible nodes are affected as a result of a malware attack.

Table 3. Comparative study with different communication radius.

		Time		
		250	500	1000
Methods	r	Counts of Infected sensor nodes		
Ojha et al. [21]	1.0	280	240	180
Awasthi et al. [3]	1.0	3.8	10	25
Proposed method	1.0	3.8	10	22
Ojha et al. [21]	1.5	280	240	180
Awasthi et al. [3]	1.5	13	50	77
Proposed method	1.5	7.2	31	77

Case 2: In this case, we assume the parameters as: $b = 0.4, \beta = 0.001, \lambda_1 = 0.002, \lambda_2 = 0.001, \sigma = 0.001, \gamma = 0.003, \omega = 0.004, p = 0.7, N = 1000, r = 1.61$, and we only change the value of distributed node density as $\varphi = 11, L = 10$ and $\varphi = 16, L = 8$. Table 4 provides the comparison between existing and proposed models in this scenario. Table 4 shows that in the proposed strategy, fewer susceptible nodes are affected by malware attacks than in previous models.

Table 4. Comparative study with different distributed node density φ.

		Time		
		250	500	1000
Methods	φ	Counts of Infected sensor nodes		
Ojha et al. [21]	11	225	160	100
Awasthi et al. [3]	11	4.7	13.3	23
Proposed method	11	4.7	13.3	23
Ojha et al. [21]	16	225	160	100
Awasthi et al. [3]	16	13	22	25
Proposed method	16	3.2	7.1	10

Thus, it can be inferred from the aforementioned analysis that the design parameters, including the sensor node's communication radius, distributed node

density, and so on, must be determined following the requirements. Indeed, considering the additional needs of the sensor nodes that will be placed in the designated area is also crucial.

7 Conclusion

This work illustrated the propagation dynamics of multi-malware infections in WSNs using epidemiological modeling. The ABM approach was employed to confirm the theoretical findings and identify the problem of malware attacks. This study investigated the efficacy of the parameters and explored how various model parameters affect the propagation of malware in WSNs. The proposed model has given significant insight into the dynamics of multi-malware propagation in WSNs and provides a technique to control and prevent malware spread. The results of employing local and global sensitivity analysis were proposed. The study further provided a comparative analysis of the proposed model with existing ones.

Acknowledgements. This work was partially supported by SERICS (PE00000014) under PNRR MUR program funded by EU-NGEU.

References

1. Acarali, D., Rajarajan, M., Komninos, N., Zarpelão, B.B.: Modelling the spread of botnet malware in IoT-based wireless sensor networks. Secur. Commun. Netw. **2019**(1), 3745619 (2019). https://doi.org/10.1155/2019/3745619
2. AL-Mousawi, A.J., AL-Hassani, H.K.: A survey in wireless sensor network for explosives detection. Comput. Electric. Eng. **72**, 682–701 (2018). https://doi.org/10.1016/j.compeleceng.2017.11.013
3. Awasthi, S., et al.: An epidemic model for the investigation of multi-malware attack in wireless sensor network. IET Commun. **17**(11), 1274–1287 (2023). https://doi.org/10.1049/cmu2.12622
4. Chernikova, A., Gozzi, N., Perra, N., Boboila, S., Eliassi-Rad, T., Oprea, A.: Modeling self-propagating malware with epidemiological models. Appl. Netw. Sci. **8**(1), 52 (2023)
5. Conte, D., Farsimadan, E., Moradi, L., Palmieri, F., Paternoster, B.: Time-delay fractional optimal control problems: a survey based on methodology. In: Abdel Wahab, M. (ed.) FFW 2020 2020. LNME, pp. 325–337. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9893-7_23
6. Conte, D., Farsimadan, E., Moradi, L., Palmieri, F., Paternoster, B.: A Galerkin approach for fractional delay differential equations using hybrid Chelyshkov basis functions. In: Gervasi, O. (ed.) Computational Science and Its Applications – ICCSA 2022, pp. 130–142. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-10522-7_10
7. Conte, D., Farsimadan, E., Moradi, L., Palmieri, F., Paternoster, B.: Numerical solution of delay Volterra functional integral equations with variable bounds. Appl. Numer. Math. **200**, 148–161 (2024). https://doi.org/10.1016/j.apnum.2023.05.002

8. De Santis, A., Farsimadan, E., Moradi, L., Palmieri, F.: Generalized defensive modeling of fake news propagation in social networks using fractional differential equations. IEEE Trans. Comput. Soc. Syst. **12**(2), 622–634 (2025). https://doi.org/10.1109/TCSS.2024.3492097
9. Farjamnia, G., Gasimov, Y., Kazimov, C.: Review of the techniques against the wormhole attacks on wireless sensor networks. Wireless Pers. Commun. **105**(4), 1561–1584 (2019)
10. Farsimadan, E., Moradi, L., Conte, D., Paternoster, B., Palmieri, F.: Comparison Between protein-protein interaction networks $CD4^+T$ and $CD8^+T$ and a numerical approach for fractional HIV infection of $CD4^+T$ cells. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12949, pp. 78–94. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86653-2_6
11. Farsimadan, E., Moradi, L., Palmieri, F.: A general study on the malware propagation models in wireless sensor networks. In: Gervasi, O. (ed.) Computational Science and Its Applications – ICCSA 2024 Workshops, pp. 83–99. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-65223-3_6
12. Frutos-Bernal, E., Rodríguez-Rosa, M., Anciones-Polo, M., Martín-del Rey, A.: Analyzing malware propagation on wireless sensor networks: a new approach using queueing theory and HJ-BIPLOT with a sirs model. Mathematics **12**(1) (2024). https://doi.org/10.3390/math12010135
13. Geetha, R., Madhusudanan, V., Srinivas, M.N.: Influence of clamor on the transmission of worms in remote sensor network. Wireless Pers. Commun. **118**(1), 461–473 (2021)
14. Giri, P., Ng, K., Phillips, W.: Wireless sensor network system for landslide monitoring and warning. IEEE Trans. Instrum. Meas. **68**(4), 1210–1220 (2019). https://doi.org/10.1109/TIM.2018.2861999
15. Han, L., Zhou, M., Jia, W., Dalil, Z., Xu, X.: Intrusion detection model of wireless sensor networks based on game theory and an autoregressive model. Inf. Sci. **476**, 491–504 (2019). https://doi.org/10.1016/j.ins.2018.06.017
16. Jamshidi, M., Zangeneh, E., Esnaashari, M., Darwesh, A.M., Meybodi, M.R.: A novel model of Sybil attack in cluster-based wireless sensor networks and propose a distributed algorithm to defend it. Wireless Pers. Commun. **105**(1), 145–173 (2019)
17. Liu, G., Chen, J., Liang, Z., Peng, Z., Li, J.: Dynamical analysis and optimal control for a seir model based on virus mutation in WSNs. Mathematics **9**(9) (2021). https://doi.org/10.3390/math9090929
18. Moradi, L., Farsimadan, E., D'Angelo, G., Carpentieri, B., Palmieri, F.: Epidemic analysis of the propagation of multiple malware infectious in wireless sensor networks. In: Barolli, L. (ed.) Advanced Information Networking and Applications, pp. 140–148. Cham (2025). https://doi.org/10.1007/978-3-031-87784-1_13
19. Ndii, M.Z., Djahi, B.S., Rumlaklak, N.D., Supriatna, A.K.: Determining the important parameters of mathematical models of the propagation of malware. In: Othman, M.A., Abd Aziz, M.Z.A., Md Saat, M.S., Misran, M.H. (eds.) Proceedings of the 3rd International Symposium of Information and Internet Technology (SYMINTECH 2018), pp. 1–9. Springer International Publishing, Cham (2019)
20. Nwokoye, C., Umeh, I.: Analytic-agent cyber dynamical systems analysis and design method for modeling spatio-temporal factors of malware propagation in wireless sensor networks. MethodsX **5**, 1373–1398 (2018). https://doi.org/10.1016/j.mex.2018.10.005

21. Ojha, R.P., Srivastava, P.K., Sanyal, G.: Mathematical model for wireless sensor network with two latent periods. In: Lobiyal, D.K., Mansotra, V., Singh, U. (eds.) Next-Generation Networks, pp. 497–504. Springer, Singapore (2018)
22. Ojha, R.P., Srivastava, P.K., Sanyal, G.: Improving wireless sensor networks performance through epidemic model. Int. J. Electron. **106**(6), 862–879 (2019). https://doi.org/10.1080/00207217.2019.1570563
23. Martín-del Rey, A.: A novel model for malware propagation on wireless sensor networks. Math. Biosci. Eng. **21**(3), 3967–3998 (2024). https://doi.org/10.3934/mbe.2024176
24. Roddam, A.W.: Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. Int. J. Epidemiol. **30**(1), 186 (2001). https://doi.org/10.1093/ije/30.1.186
25. Shen, S., Huang, L., Zhou, H., Yu, S., Fan, E., Cao, Q.: Multistage signaling game-based optimal detection strategies for suppressing malware diffusion in fog-cloud-based IoT networks. IEEE Internet Things J. **5**(2), 1043–1054 (2018). https://doi.org/10.1109/JIOT.2018.2795549
26. Shen, S., et al.: HSIRD: a model for characterizing dynamics of malware diffusion in heterogeneous WSNs. J. Netw. Comput. Appl. **146**, 102420 (2019). https://doi.org/10.1016/j.jnca.2019.102420
27. Shen, S., Zhou, H., Feng, S., Liu, J., Cao, Q.: SNIRD: disclosing rules of malware spread in heterogeneous wireless sensor networks. IEEE Access **7**, 92881–92892 (2019). https://doi.org/10.1109/ACCESS.2019.2927220
28. Singh, A., Awasthi, A.K., Singh, K., Srivastava, P.K.: Modeling and analysis of worm propagation in wireless sensor networks. Wireless Pers. Commun. **98**(3), 2535–2551 (2018)
29. Upadhyay, R.K., Kumari, S.: Bifurcation analysis of an e-epidemic model in wireless sensor network. Int. J. Comput. Math. **95**(9), 1775–1805 (2018). https://doi.org/10.1080/00207160.2017.1336550
30. Wu, G., Xie, L., Zhang, H., Wang, J., Shen, S., Yu, S.: STSIR: an individual-group game-based model for disclosing virus spread in social internet of things. J. Netw. Comput. Appl. **214**, 103608 (2023). https://doi.org/10.1016/j.jnca.2023.103608
31. Wu, J., Jiang, W., Mei, Y., Zhou, Y., Wang, T.: A survey on the progress of testing techniques and methods for wireless sensor networks. IEEE Access **7**, 4302–4316 (2019). https://doi.org/10.1109/ACCESS.2018.2887246
32. Wu, X., Cao, Q., Jin, J., Li, Y., Zhang, H.: Nodes availability analysis of NB-IoT based heterogeneous wireless sensor networks under malware infection. Wirel. Commun. Mob. Comput. **2019**(1), 4392839 (2019). https://doi.org/10.1155/2019/4392839
33. Xu, B., Lu, M., Zhang, H., Pan, C.: A novel multi-agent model for robustness with component failure and malware propagation in wireless sensor networks. Sensors **21**(14) (2021). https://doi.org/10.3390/s21144873
34. Ye, X., Xie, S., Shen, S.: SIR1R2: characterizing malware propagation in WSNs with second immunization. IEEE Access **9**, 82083–82093 (2021). https://doi.org/10.1109/ACCESS.2021.3086531
35. Yuan, Y., Huo, L., Wang, Z., Hogrefe, D.: Secure APIT localization scheme against sybil attacks in distributed wireless sensor networks. IEEE Access **6**, 27629–27636 (2018). https://doi.org/10.1109/ACCESS.2018.2836898
36. Zhang, H., Yang, S., Wu, G., Shen, S., Cao, Q.: Steady-state availability evaluation for heterogeneous edge computing-enabled WSNs with malware infections. Mob. Inf. Syst. **2022**(1), 4743605 (2022). https://doi.org/10.1155/2022/4743605

37. Zhang, Q., Zhang, W.: Accurate detection of selective forwarding attack in wireless sensor networks. Int. J. Distrib. Sens. Netw. **15**(1), 1550147718824008 (2019). https://doi.org/10.1177/1550147718824008
38. Zhou, H., Shen, S., Liu, J.: Malware propagation model in wireless sensor networks under attack–defense confrontation. Comput. Commun. **162**, 51–58 (2020). https://doi.org/10.1016/j.comcom.2020.08.009
39. Zhou, Y., Wang, Y., Zhou, K., Shen, S.F., Ma, W.X.: Dynamical behaviors of an epidemic model for malware propagation in wireless sensor networks. Front. Phys. **11** (2023). https://doi.org/10.3389/fphy.2023.1198410

Computational Astrochemistry 2026 (CompAstro 2026)

From N(^{2}D) Back to N_2: A Theoretical Investigation of the N(^{2}D) + $CH_2NH \rightarrow N_2 + CH_3$ Reaction and Implications for the Atmospheric Chemistry of Titan

Luca Mancini[1(✉)], Marco Parriani[1], Marzio Rosi[2], and Nadia Balucani[1]

[1] Dipartimento di Chimica, Biologia e Biotecnologie, Università degli Studi di Perugia, 06123 Perugia, Italy
luca.mancini@unipg.it

[2] Dipartimento di Ingegneria Civile ed Ambientale, Università degli Studi di Perugia, 06125 Perugia, Italy

Abstract. Nitrogen atoms in their first electronically excited state, ^{2}D, are key reactive species in Titan's upper atmosphere, playing a crucial role in the formation of nitrogen-rich organic molecules and shaping the complex organic chemistry of this environment. In this work, we present a theoretical investigation of the reaction N(^{2}D) + $CH_2NH \rightarrow N_2 + CH_3$. Using a combination of density functional theory (DFT) and high-level Coupled-Cluster methods, we explored the stationary points along the minimum energy pathway. Our results indicate that the reaction proceeds without an entrance barrier, with all the internal rearrangements characterized by modest barriers lying well below the reactant energy. Therefore, the reaction is feasible under the conditions of Titan. It remains to be seen whether competitive reaction channels leading to other product sets are open. Given that the actual abundance of CH_2NH is uncertain and may be overestimated in current photochemical models, our findings provide an additional route of methanimine destruction that may help reconcile the predictions of the model with the observed nitrogen chemistry.

Keywords: DFT · Titan's Atmosphere · Potential Energy Surface · Reaction Dynamics

1 Introduction

Methanimine (CH_2NH) is considered a key molecule in prebiotic chemistry [1] due to its potential role as a precursor to glycine being a crucial intermediate in the Strecker synthesis of aminoacetonitrile [2,3]. The presence of methanimine in the upper atmosphere of Titan, the largest moon of Saturn, was established through the analysis of ion spectra obtained by the Cassini Ion Neutral Mass Spectrometer (INMS) which revealed a significant peak at the mass-to-charge ratio of 30 attributed to protonated methanimine [4], with a mole fraction of approximately 1×10^{-5} for the neutral form at an altitude of 1100 km [5]. Since its identification, a number of elementary processes that may

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 425–435, 2027.
https://doi.org/10.1007/978-3-032-30521-3_26

occur in the thermosphere of Titan have been proposed as possible formation routes and studied using laboratory experiments and/or quantum-chemical calculations [6–9]. Among the considered processes are *i*) reactions between neutral species involving the first excited electronic state of atomic nitrogen (^{2}D) and the abundant CH_4 and C_2H_6 [6,7] and the CH + NH_3 [8] or NH + CH_3/C_2H_5 [10–12] reactions, *ii*) the photodissociation of methylamine, and *iii*) some ion-molecule reactions including electron-ion dissociative recombination of $CH_2NH_2^+$ [9].

Despite its importance, significant discrepancies exist between the predicted and observed mixing ratios of methanimine in Titan's atmosphere, with the predicted abundance consistently overestimating the measured value in all available photochemical models [11,13–15]. This is likely due to limited knowledge of its chemical behavior and, especially, loss mechanisms. Methanimine is a highly reactive molecule that polymerizes readily under terrestrial conditions unless maintained at very low temperatures. This property has led to speculation that its polymerization could contribute to the formation of nitrogen-rich organic aerosols [14,15], which are believed to contribute to Titan's characteristic haze. However, previous studies suggest that the dimerization of neutral methanimine molecules in the gas phase is hindered by significant energy barriers, making it unlikely under the cold conditions of Titan's atmosphere or interstellar space [16]. On the contrary, the processes involving a single molecule of either ionized or protonated methanimine are barrierless and can therefore contribute to the formation of larger ions. In particular, the reaction involving ionized methanimine and neutral methanimine can also be efficient in the very low temperature conditions of Titan, leading to products of general formula $C_2N_2H_5^+$ [17]. The reactivity of methanimine radical cation (H_2CNH^+) and its isomer aminomethylene ($HCNH_2^+$) with the most abundant hydrocarbons of Titan was also investigated [18–20]. Finally, the search for other reactions that can consume methanimine in Titan's atmosphere included the reaction with the CN radical [21], which we know to be abundant, but reactions with C_2H and C_3N are also possible (see [22,23]).

More recently, we have turned our attention to a possible mechanism for N_2 regeneration from the reactions of N(^{2}D) with nitriles (or other nitrogen-containing organic species). This interest stems from the fact that the main channel for the N(^{2}D) + HCN reaction is believed to be the one leading to N_2 + CH [24]. For this reason, we searched for similar channels for the N(^{2}D) reactions with acrylonitrile, acetonitrile, and cyanoacetylene [25–27]. In all cases, the channels leading to N_2 + C_3H_3, CH_2CH/CH_3C, l-C_3H are open, but not competitive. Somewhat surprisingly, we also found that the reaction with pyridine leads to the formation of N_2 and C_5H_5 (cyclopentadienyl), albeit in very modest yield [28]. In this work, we intend to analyze the case of the N(^{2}D) + CH_2NH reaction. The nitrogen atom of methanimine is not involved in a triple bond, as in the case of nitriles. Therefore, it will be very interesting to verify whether the reaction path leading to N_2 + CH_3 is feasible under the conditions of the thermosphere of Titan.

2 Theoretical Methods

The Potential Energy Surface (PES) for the reaction N(^{2}D) + CH_2NH was investigated combining DFT calculations and a more accurate coupled-cluster analysis. As

already done in several cases [26,29–33], geometry optimization and harmonic vibrational frequencies calculations were performed through density functional theory, with the B3LYP functional [34,35] combined with the correlation consistent valence polarized set aug-cc-pVTZ basis set [36–38]. Additionally, coupled-cluster [CCSD(T)] [39–41] calculations were performed, with the same basis set, to obtain more accurate values of energy to be used in the PES. All the energy values were corrected using the zero-point energy, evaluated at the DFT level of theory. All calculations were performed using the Gaussan09 software [42] and the visualization of molecular structure was performed using Avogadro [43]. As already noticed in previous investigations [26], single-reference methods, such as B3LYP, provide an unsatisfactory description of the lowest spin excited state. The same results can be obtained including correlation effects (e.g. CCSD(T) calculations). For this reason, the energy of N(^{2}D) was estimated considering the experimental energy difference to N(^{4}S), which corresponds to 230.0 kJ/mol [44]. This value was added to the energy of N(^{4}S) at all levels of the calculation.

3 Results

As a first step of our analysis, a detailed investigation of the entrance channel was performed, in order to identify the presence of a barrier related to the attack of N(^{2}D) to the methanimine molecule. For this reason, a *scan* calculation was carried out at the DFT level, increasing the C-N distance up to 6 Å, with a step of 0.1 Å end evaluating the energy at each step. The resulting plot of the energy as a function of the distance is reported in Fig. 1.

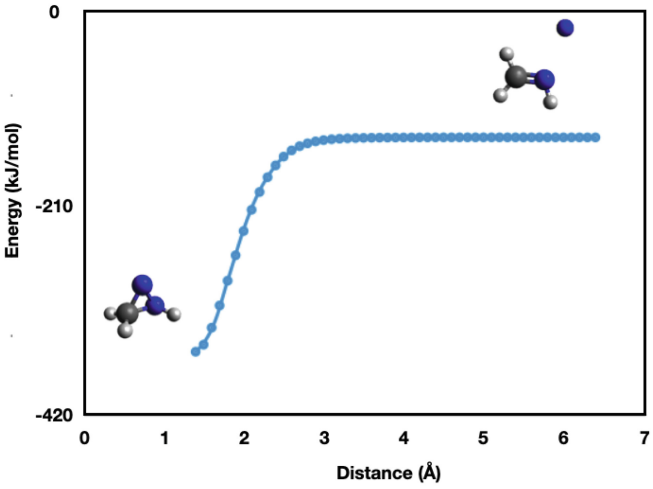


Fig. 1. Energy (kJ/mol) as a function of the C-N distance evaluated for the initial attack of N(^{2}D) to the CH_2NH molecule.

As shown, the energy profile as a function of the intermolecular distance decreases monotonically with decreasing C-N distance, with no evidence of a local maximum along the entrance channel. This behavior indicates the absence of an initial energy barrier, as expected, suggesting that the interaction is barrierless along the explored coordinate.

The potential energy surface is reported in Fig. 2.

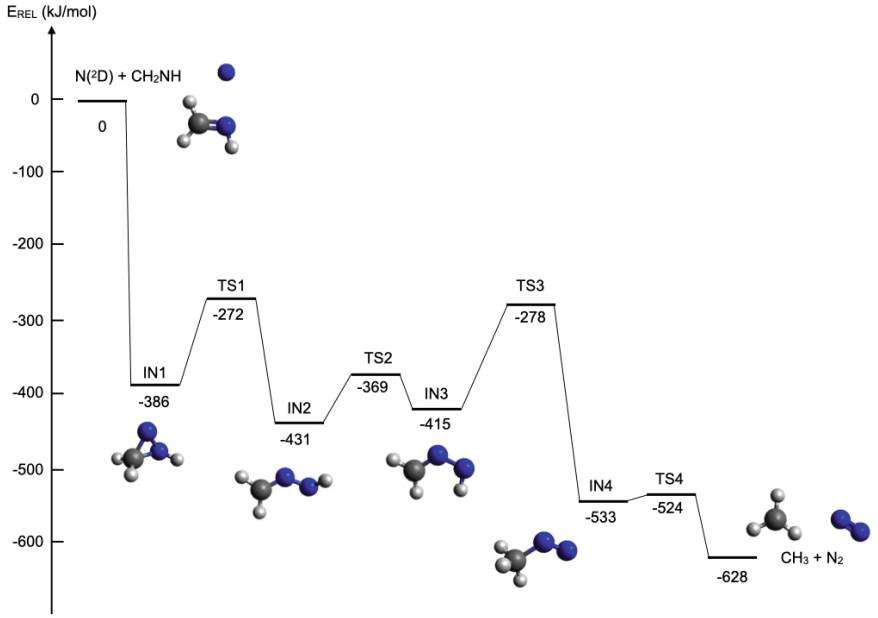


Fig. 2. Schematic representation of the potential energy surface obtained for the $N(^2D)+CH_2NH$ reaction, evaluated at the CCSD(T)/aug-cc-pVTZ//B3LYP/aug-cc-pVTZ level of theory.

The first step of the reaction is the barrierless attack of $N(^2D)$ to the C-N double bond of the methanimine molecule, leading to the formation of a stable intermediate (IN1), located 386 kJ/mol below the reactant energy asymptote and characterized by the presence of a three-membered ring in the molecular structure. A barrier of 114 kJ/mol is associated with the C-N bond fission, leading to the formation of a linear intermediate, characterized by the presence of a C-N-N chain. Two different isomers were identified (IN2 and IN3), with a relative energy of −431 and −415 kJ/mol respectively, and connected by a small barrier of 62 kJ/mol (represented by TS2). A C-N hydrogen transfer can be achieved, forming the CH_3NN intermediate (named IN4 in the PES), which

shows a relative energy of −533 kJ/mol. The related transition state (TS3) is located 278 kJ/mol below the energy of the reactants and clearly shows the transfer of an H atom from a nitrogen to the carbon atom of the structure. Finally, a small barrier of 9 kJ/mol is associated to the C-N bond breaking and leads to the formation of molecular nitrogen (N_2) and methyl radical (CH_3). The exothermicity of the products is 628 kJ/mol. In Figs. 3, 4 and 5 are reported the geometries of the stationary points identified in the potential energy surface.

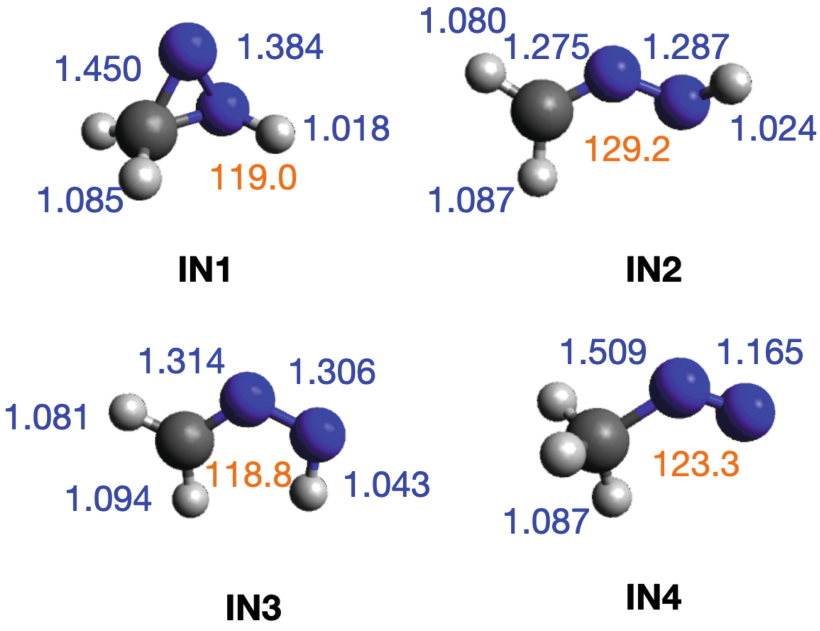


Fig. 3. B3LYP optimized geometries (distances in Å, in blue; angles in orange, in °) of the intermediates identified along the PES for the reaction N(^{2}D)+CH_2NH. (Color figure online)

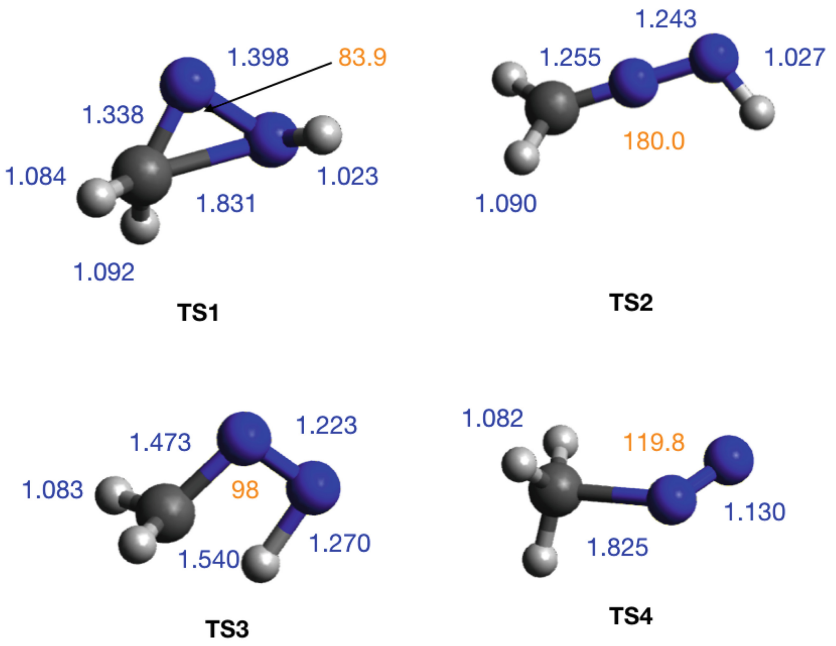


Fig. 4. B3LYP optimized geometries (distances in Å, in blue; angles in orange, in °) of the transition states identified along the PES for the reaction $N(^2D)+CH_2NH$. (Color figure online)

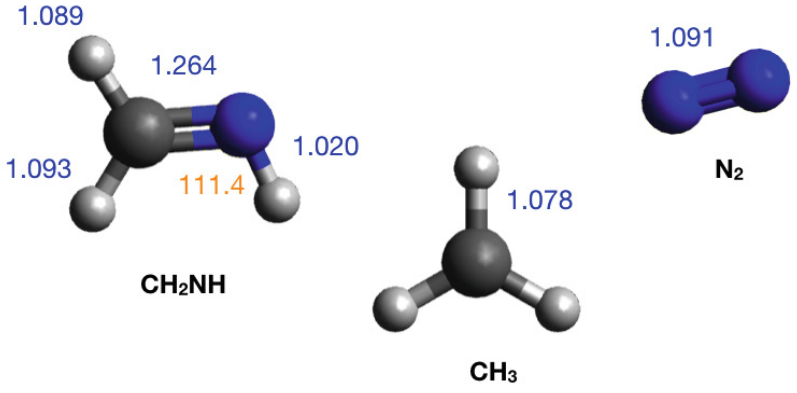


Fig. 5. B3LYP optimized geometries (distances in Å, in blue; angles in orange, in °) of the reactant and products identified along the PES for the reaction $N(^2D)+CH_2NH$ (Color figure online)

4 Discussion

The present results can be compared with a series of previous investigation performed on the reaction of N(^{2}D) with different nitrile species, including acetonitrile (CH_3CN), vinyl cyanide (CH_2CHCN) and cyanoacetylene (HC_3N) [25–27], where the same computational approach was used for the analysis of the PES. In the case of CH_3CN, CH_2CHCN and HC_3N the presence of a submerged barrier along the nitrogen-attack pathway (ranging from −5 to −10 kJ/mol) effectively reduces its competitiveness with respect to alternative addition channels. This behavior can be rationalized by considering the electronic structure of the –CN moiety in the three molecules, where the strong π-bonding orbitals favor attack on the carbon side, while the lone pair on the nitrile nitrogen is stabilized and less available for interaction with the electrophilic N(^{2}D). A different case is represented by the CH_2NH molecule, where the C = N bond is less stabilized than the triple bond in nitriles. Consequently, the reaction proceeds in a barrierless approach toward the formation of a three-membered C–(N)–N cyclic intermediate. This intermediate reflects efficient overlap between the singly occupied orbital of N(^{2}D) and the lone-pair/π system of the imine, without the need to overcome an entrance barrier. The minimum-energy path of the reaction N(^{2}D) + $CH_2NH \rightarrow N_2$ + CH_3 is characterized by local maxima and minima that lie at energies well below the asymptotic energy of the reactants. There is therefore no doubt that this reaction channel is open under the conditions of Titan's thermosphere, where methanimine reaches its maximum mixing ratio, and the N(^{2}D) production rate is at its maximum.

The impact of temperature on the N(^{2}D) + CH_2NH reaction under Titan's thermospheric conditions is expected to be moderate, although not negligible. Since the N(^{2}D) electronic state is highly reactive and the reaction with methanimine is expected to proceed without a significant entrance barrier, the overall rate coefficient should show only a weak temperature dependence, as is typical for barrierless radical-neutral reactions.

Temperature may still affect the reaction by influencing the branching ratios among competing product channels, by modifying the stabilization and lifetime of intermediate complexes, particularly at lower temperatures. Nevertheless, as previously pointed out, these effects are expected to remain limited over the relatively narrow temperature range of Titan's temperature.

In this contribution, we have not explored competing reaction channels that, however, could be present if we consider the N(^{2}D) addition to the C-atom of methanimine. Therefore, to draw definitive conclusions about the importance of this reaction pathway, we need to complete the study and determine the branching ratios for the competing pathways. In all cases, the characterized pathway remains facile and perfectly available under the conditions of Titan's thermosphere.

5 Conclusions

In the present contribution, a theoretical investigation of a possible path in the reaction of N(^{2}D) with methanimine (CH_2NH) was performed combining DFT theory and CCSD(T) approach to characterize the full potential energy surface. The calculations show how the initial attack of N(^{2}D) on the C = N bond of methanimine is barrierless,

in contrast with similar reactions with nitriles such as acetonitrile, vinyl cyanide, and cyanoacetylene, where submerged barriers along the nitrogen-attack pathway reduce the competitiveness of this channel. Subsequent steps leading to the formation of CH_3 + N_2 involve intermediates and transition states that lie significantly below the energy of the separated reactants, confirming that the overall process is exothermic and barrierless with respect to the entrance channel. These findings indicate that the $N(^2D)$ + $CH_2NH \rightarrow CH_3$ + N_2 reaction can be a feasible and efficient pathway under the conditions of Titan's thermosphere, where both reactants are present and abundant. While other potential reaction pathways, characterized by different rearrangements of the initial intermediate, may compete, the present findings show a plausible route for the formation of N_2 in Titan's atmosphere. This work provides a foundation for future theoretical studies, aimed at a detailed analysis of the complete PES, with a focus on quantifying product branching ratios to fully assess the astrochemical impact of this process.

Acknowledgments. Support from the Italian Space Agency (Bando ASI Prot. n. DC-DSR-UVS-2022-231, Grant no. 2023-10-U.0 MIGLIORA) is acknowledged. MR acknowledges financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title 2022JC2Y93 ChemicalOrigins: linking the fossil composition of the Solar System with the chemistry of protoplanetary disks – CUP J53D23001600006 - Grant Assignment Decree No. 962 adopted on 30.06.2023 by the Italian Ministry of Ministry of University and Research (MUR). NB acknowledges additional support by the International Space Science Institute (ISSI) in Bern, through ISSI International Team project 615 "Advancing Titan's Atmospheric Chemistry Knowledge". LM and MR acknowledge the Project STEAMigPOWER: STEAM approaches at higher education for mIGrants, refugees and asylum seekers' emPOWERment under the ERASMUS+ Program key action 2: Cooperation partnership in higher education – Agreement n. 2022–1-PT01-KA220-HED-000088221.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Balucani, N.: Elementary reactions of N atoms with hydrocarbons: first steps towards the formation of prebiotic N-containing molecules in planetary atmospheres. Chem. Soc. Rev. **41**(16), 5473–5483 (2012)
2. Danger, G., et al.: Experimental investigation of aminoacetonitrile formation through the Strecker synthesis in astrophysical-like conditions: reactivity of methanimine (CH_2NH), ammonia (NH_3), and hydrogen cyanide (HCN). Astron. Astrophys. **535**, A47 (2011)
3. Rimola, A., Sodupe, M., Ugliengo, P.: Deep-space glycine formation via Strecker-type reactions activated by ice water dust mantles. A computational approach. Phys. Chem. Chem. Phys. **12**(20), 5285–5294 (2010)
4. Vuitton, V., Yelle, R.V., Anicich, V.G.: The nitrogen chemistry of Titan's upper atmosphere revealed **647**(2), L175–L178 (2006). https://doi.org/10.1086/507467
5. Vuitton, V., Yelle, R.V., McEwan, M.J.: Ion chemistry and N-containing molecules in Titan's upper atmosphere **191**(2), 722–742 (2007). https://doi.org/10.1016/j.icarus.2007.06.023

6. Balucani, N., et al.: Combined crossed molecular beam and theoretical studies of the N(^{2}D)+ CH_4 reaction and implications for atmospheric models of Titan. J. Phys. Chem. A **113**(42), 11138–11152 (2009)
7. Balucani, N., et al.: Formation of nitriles and imines in the atmosphere of Titan: combined crossed-beam and theoretical studies on the reaction dynamics of excited nitrogen atoms N(^{2}D) with ethane. Faraday Discuss. **147**, 189–216 (2010)
8. Yang, Z., Medvedkov, I.A., Goettl, S.J., Kaiser, R.I.: Low-temperature gas-phase formation of methanimine (CH_2NH; X^1A') – the simplest imine – under single-collision conditions. J. Phys. Chem. Lett. **14**(38), 8500–8506 (2023). https://doi.org/10.1021/acs.jpclett.3c02360
9. Yuen, C.H., Ayouz, M.A., Balucani, N., Ceccarelli, C., Schneider, I.F., Kokoouline, V.: Dissociative recombination of CH_2NH_2+: a crucial link with interstellar methanimine and titan ammonia. Mon. Not. Roy. Astron. Soc. **484**(1), 659–664 (2019). https://doi.org/10.1093/mnras/sty3514
10. Yelle, R.V., et al.: Formation of NH_3 and CH_2 NH in titan's upper atmosphere. Faraday Discuss. **147**, 31–49 (2010)
11. Vuitton, V., Yelle, R., Klippenstein, S., Hörst, S., Lavvas, P.: Simulating the density of organic species in the atmosphere of Titan with a coupled ion-neutral photochemical model. Icarus **324**, 120–197 (2019)
12. Balucani, N., Skouteris, D., Ceccarelli, C., Codella, C., Falcinelli, S., Rosi, M.: A theoretical investigation of the reaction between the amidogen, NH, and the ethyl, C_2H_5, radicals: a possible gas-phase formation route of interstellar and planetary ethanimine. Mol. Astrophys. **13**, 30–37 (2018). https://doi.org/10.1016/j.molap.2018.10.001
13. Loison, J.C., et al.: The neutral photochemistry of nitriles, amines and imines in the atmosphere of Titan **247**, 218–247 (2015). https://doi.org/10.1016/j.icarus.2014.09.039
14. Lavvas, P., Coustenis, A., Vardavas, I.: Coupling photochemistry with haze formation in titan's atmosphere, Part I: model description. Planet. Space Sci. **56**(1), 27–66 (2008)
15. Lavvas, P., Coustenis, A., Vardavas, I.: Coupling photochemistry with haze formation in Titan's atmosphere, Part II: results and validation with Cassini/Huygens data. Planet. Space Sci. **56**(1), 67–99 (2008)
16. Rosi, M., Falcinelli, S., Balucani, N., Casavecchia, P., Skouteris, D.: A theoretical study of formation routes and dimerization of methanimine and implications for the aerosols formation in the upper atmosphere of Titan. In: Murgante, B., et al. (eds.) ICCSA 2013. LNCS, vol. 7971, pp. 47–56. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39637-3_4
17. Skouteris, D., Balucani, N., Faginas-Lago, N., Falcinelli, S., Rosi, M.: Dimerization of methanimine and its charged species in the atmosphere of titan and interstellar/cometary ice analogs. Astron. Astrophys. **584**, A76 (2015)
18. Richardson, V., et al.: The reactivity of methanimine radical cation (H_2CNH^+) and its isomer aminomethylene ($HCNH_2^+$) with methane. Chem. Phys. Lett. **775**, 138611 (2021). https://doi.org/10.1016/j.cplett.2021.138611
19. Sundelin, D., et al.: The reactivity of methanimine radical cation (H_2CNH^+) and its isomer aminomethylene ($HCNH_2^+$) with C_2H_4. Chem. Phys. Lett. **777**, 138677 (2021). https://doi.org/10.1016/j.cplett.2021.138677
20. Richardson, V., et al.: Experimental and computational studies on the reactivity of methanimine radical cation (H_2CNH+•) and its isomer aminomethylene ($HCNH_2$+•) with C_2H_2. Front. Astron. Space Sci. **8**, 752376 (2021)
21. Vazart, F., Latouche, C., Skouteris, D., Balucani, N., Barone, V.: Cyanomethanimine isomers in cold interstellar clouds: insights from electronic structure and kinetic calculations. Astrophys. J. **810**(2), 111 (2015)
22. Lupi, J., Puzzarini, C., Barone, V.: Methanimine as a key precursor of imines in the interstellar medium: the case of propargylimine. Astrophys. J. Lett. **903**(2), L35 (2020)

23. Ye, H., Alessandrini, S., Puzzarini, C.: On the general mechanism for the gas-phase reaction of methanimine with a radical species in the interstellar medium: Some failures and an important success. Astrophys. J. **962**(1), 32 (2024)
24. Hébrard, E., Dobrijevic, M., Loison, J.C., Bergeat, A., Hickson, K.M.: Neutral production of hydrogen isocyanide (HNC) and hydrogen cyanide (HCN) in titans upper atmosphere. A&A **541**, A21 (2012)
25. Vanuzzo, G., et al.: The N (2D)+ CH_2CHCN (vinyl cyanide) reaction: a combined crossed molecular beam and theoretical study and implications for the atmosphere of titan. J. Phys. Chem. A **126**(36), 6110–6123 (2022)
26. Mancini, L., et al.: A computational characterization of the reaction mechanisms for the reactions N (2D)+ CH_3CN and HC_3N and implications for the nitrogen-rich organic chemistry of titan. Comput. Theor. Chem. **1229**, 114341 (2023)
27. Liang, P., et al.: Combined crossed molecular beams and computational study on the n (2D)+ HCCCN (x1σ+) reaction and implications for extra-terrestrial environments. Mol. Phys. **120**(1–2), e1948126 (2022)
28. Mancini, L., et al.: Unveiling the reaction mechanism of the N(2D) + pyridine reaction: ring-contraction versus 7-membered-ring formation channels. J. Phys. Chem. A **128**(34), 7177–7194 (2024)
29. Liang, P., et al.: Oh (2π)+ C_2H_4 reaction: a combined crossed molecular beam and theoretical study. J. Phys. Chem. A **127**(21), 4609–4623 (2023)
30. Rosi, M., et al.: Electronic structure and kinetics calculations for the SI+ SH reaction, a possible route of sis formation in star-forming regions. In: International Conference on Computational Science and Its Applications, pp. 306–315. Springer (2019)
31. Mancini, L., et al.: Destruction of interstellar methyl cyanide (CH_3CN) via collisions with he+ ions. Astron. Astrophys. **691**, A83 (2024)
32. Richardson, V., et al.: Fragmentation of interstellar methanol by collisions with He+: an experimental and computational study. Phys. Chem. Chem. Phys. **24**(37), 22437–22452 (2022)
33. de Aragão, V.F.E., Mancini, L., Faginas-Lago, N., Rosi, M., Skouteris, D., Pirani, F.: Semiempirical potential in kinetics calculations on the HC_3N+ CN reaction. Molecules **27**(7), 2297 (2022)
34. Becke, A.D.: A new mixing of Hartree-Fock and local density-functional theories. J. Chem. Phys. **98**(2), 1372–1377 (1993)
35. Stephens, P.J., Devlin, F.J., Chabalowski, C.F., Frisch, M.J.: Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. J. Phys. Chem. **98**(45), 11623–11627 (1994)
36. Dunning, Jr. T.H.: Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. J. Chem. Phys. **90**(2), 1007–1023 (1989)
37. Woon, D.E., Dunning, Jr. T.H.: Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. J. Chem. Phys. **98**(2), 1358–1371 (1993)
38. Kendall, R.A., Dunning, Jr. T.H., Harrison, R.J.: Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. J. Chem. Phys. **96**(9), 6796–6806 (1992)
39. Bartlett, R.J.: Many-body perturbation theory and coupled cluster theory for electron correlation in molecules. Annu. Rev. Phys. Chem. **32**(1), 359–401 (1981)
40. Raghavachari, K., Trucks, G.W., Pople, J.A., Head-Gordon, M.: A fifth-order perturbation comparison of electron correlation theories. Chem. Phys. Lett. **157**(6), 479–483 (1989)
41. Olsen, J., Jo/rgensen, P., Koch, H., Balkova, A., Bartlett, R.J.: Full configuration–interaction and state of the art correlation calculations on water in a valence double-zeta basis with polarization functions. J. Chem. Phys. **104**(20), 8007–8015 (1996)
42. Frisch, M., et al.: Fox: Gaussian 09, Revision A. 02, 2009. Gaussian. Inc., Wallingford CT (2009)

43. Hanwell, M.D., Curtis, D.E., Lonie, D.C., Vandermeersch, T., Zurek, E., Hutchison, G.R.: Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J. Cheminf. **4**(1), 17 (2012)
44. Schofield, K.: Critically evaluated rate constants for gaseous reactions of several electronically excited species. J. Phys. Chem. Ref. Data **8**(3), 723–798 (1979)

Production and Characterization of Cyanoacetylene Dications by Synchrotron Radiation

Marco Parriani[1,2], Franco Vecchiocattivi[1], Fernando Pirani[1,2], Michele Alagia[3], Robert Richter[4], Stefano Stranges[3,5], Raimund Feifel[6], Simonetta Cavalli[2], and Stefano Falcinelli[1(✉)]

[1] Department of Civil and Environmental Engineering, University of Perugia, Via G. Duranti 93, 06125 Perugia, Italy
franco@vecchio.it, stefano.falcinelli@unipg.it
[2] Department of Chemistry, Biology and Biotechnologies, University of Perugia, Via Elce di Sotto 8, 06100 Perugia, Italy
simonetta.cavalli@unipg.it
[3] IOM CNR Laboratorio TASC, 34012 Trieste, Italy
alagia@iom.cnr.it, stefano.stranges@uniroma1.it
[4] Sincrotrone Trieste, Area Science Park, Basovizza, 34149 Trieste, Italy
robert.richter@elettra.trieste.it
[5] Department of Chemistry and Drug Technology, University of Rome Sapienza, 00185 Rome, Italy
[6] Department of Physics, University of Gothenburg, Origovägen 6B, 412 58 Gothenburg, Sweden
raimund.feifel@physics.gu.se

Abstract. In this work, the production of the cyanoacetylene dication and the characterization of its fragmentation dynamics followed by Coulomb explosion are presented, as well as its possible role in astrochemistry. The molecular dication $(HC_3N)^{2+}$ was formed by double photoionization of the neutral precursor molecule HC_3N after its in-situ synthesis. The dissociation dynamics has been studied using the photoelectron-photoion-photoion coincidence (PEPIPICO) technique coupled with time-of-flight (TOF) mass spectrometry and synchrotron radiation in the photon energy range of 30.0–50.0 eV. Preliminary results, analyzed by a Monte Carlo computational approach, are presented regarding the threshold energy for the formation of the molecular dication $(HC_3N)^{2+}$ and of all the open fragmentation channels resulting from its Coulomb explosion. Furthermore, a more in-depth analysis of the experimental data is outlined that can determine: i) the relative cross sections of the observed fragmentation channels as a function of photon energy; ii) the kinetic energy released for the formation of each fragment ion. This work is in progress in our laboratory and highlights its relevance for fully characterizing the energetics and microscopic dynamics of a process important for both environmental chemistry and astrochemistry.

Keywords: Astrochemistry · double photoionization · synchrotron radiation · molecular dications · cyanoacetylene

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 436–447, 2027.
https://doi.org/10.1007/978-3-032-30521-3_27

1 Introduction

Despite the importance of the cyanoacetylene molecule, mainly due to its great astrochemical interest and its fundamental nature as a relatively small and reactive species, experimental data regarding the energetics and dynamics of fragmentation are still scarce.

Indeed, cyanoacetylene (HC_3N), see Fig. 1, is an important nitrogen-containing product that can be formed by the action of an electrical discharge on a mixture of methane and nitrogen. It reacts with simple inorganic substances in aqueous solution to give products including aspartic acid, asparagine, and cytosine. Furthermore, cytosine can be formed by the reaction between cyanoacetylene and cyanate, which in prebiotic chemistry is considered a potential route to the production of RNA bases on early Earth [1]. The first precursor for this reaction, an abundant interstellar molecule, is produced by an electrical discharge in a CH_4/N_2 mixture [1, 2]. However, this molecule is not commercially available and for this reason its experimental characterization, mainly from a spectroscopic point of view, is challenging.

On the other hand, the importance of ions in space has been highlighted in several papers [3–6] where the role of multiply charged ions is well emphasized. In this context, particular attention must be paid to molecular doubly positive charged ions, commonly called molecular dications: they are detected in many natural plasmas as comet tails, interstellar clouds and planetary ionospheres [7, 8]. This is the case, for example, of N_2^{2+}, O_2^{2+}, CO_2^{2+} which can be formed by double ionization induced by cosmic rays and UV-EUV photons in the upper atmospheres of Venus, Mars and Titan, Saturn's largest moon [9, 10]. Once formed, these molecular dications are in a metastable electronic state and, through a Coulomb explosion process, dissociate producing fragment monocations with a kinetic energy of several eV which allows some of them to overcome the gravitational attraction with the planet and disperse into space. [11, 12]. Several recent studies have demonstrated that these processes are responsible for the slow erosion that some planetary atmospheres, such as that of Mars, undergo due to ion escape. The latter can be generated, as mentioned above, by the Coulomb explosion of the metastable dication when it forms due to double ionization phenomena affecting the neutral precursor molecule in the ionosphere of these planets [13–15].

It is therefore important to carefully consider that the role of metastable molecular dications in the production of fast fragment ions capable of escaping into space is still underestimated in the kinetic modeling of the physical chemistry of planetary ionospheres. The goal of this type of study is also to stimulate research in the field of chemical and physical characterization of planetary atmospheres, in order to fully and more carefully consider the possible formation of molecular dications for more accurate modeling of what happens in such environments.

2 Experimental

The experiment was performed during a recent beamtime approved and performed at the GasPhase beamline of the ELETTRA Synchrotron Facility of Basovizza, Trieste, Italy (Proposal-Nr. 20235063). We were able to successfully perform the synthesis of cyanoacetylene through a procedure already well tested by our group following the one

proposed by Miller and Lemmon [16] where cyanoacetylene can be obtained at 225 °C by dehydrating C_3H_3NO (propiolamide), which has been purchased from Merck, using P_2O_5, as follows:

$$C_3H_3NO + P_2O_5 \rightarrow C_3HN + P_2O_5 - H_2O \rightarrow HC_3N + H_3PO_4 \quad (1)$$

The reaction (1) was carried out using a Pyrex system schematized in Fig. 1: round-bottom flasks with three necks are connected to a vacuum pump and a Pyrex glass container immersed in liquid nitrogen.

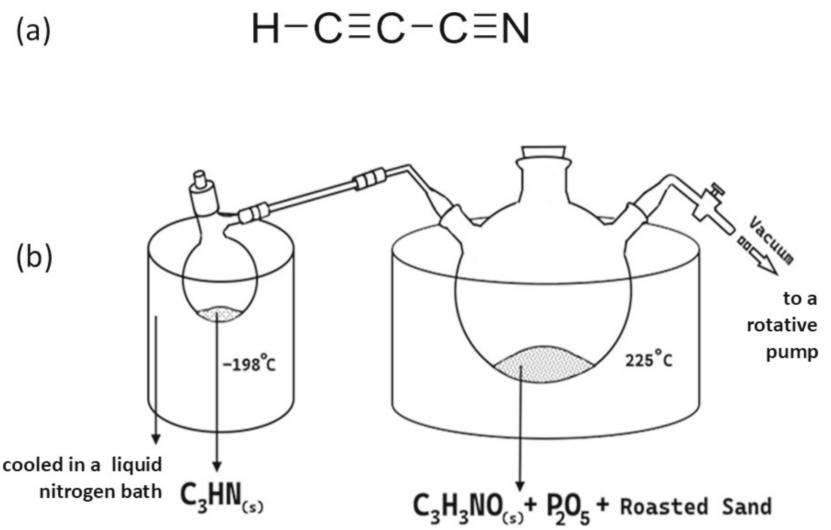


Fig. 1. (a) The Lewis structure of the cyanoacetylene molecule. (b) The scheme of the Pyrex glass setup used for the synthesis of cyanoacetylene.

The entire set-up was previously evacuated in order to remove any humidity present through the use of a rotary vacuum pump capable of obtaining a residual pressure of 10^{-2}–10^{-3} mbar. Further details of this synthesis procedure are reported in a recent article, to which we refer interested readers [17].

Then, a high-grade purity sample of cyanoacetylene has been used to produce an effusive beam exploiting its typical partial pressure of 100 mbar at room temperature. This allowed to accurately measure the threshold energy for the formation of the cyanoacetylene $(HC_3N)^{2+}$ molecular dication and for the various fragmentation channels produced by double photoionization of HC_3N in the photon energy range between 30.0 and 50.0 eV with a 100 meV energy step at the GasPhase beamline of Elettra Synchrotron Facility (Trieste, Italy).

Furthermore, by analyzing the coincidence plots recorded at each studied photon energy using the photoelectron-photoion-photoion coincidence (PEPIPICO) device mounted in the Angle-Resolved PhotoEmission Spectroscopy (ARPES) end station [18] with the use of time-of-flight (TOF) mass spectrometry, we were able to extract the threshold energy for all accessible dissociation channels. Furthermore, through careful

data analysis work that is still in progress, we will also be able to obtain the kinetic energy released (KER) distributions for the formation of each fragment ion. The ARPES station of the GasPhase beamline, whose picture is reported in Fig. 2, is a molecular beam apparatus operating in a typical vacuum regime between 10^{-7} and 10^{-8} mbar, which allow photoionization studies to be performed using a tunable and high photon flux of a third-generation synchrotron source as it is the Elettra Synchrotron Facility of Trieste, Italy [17, 18].

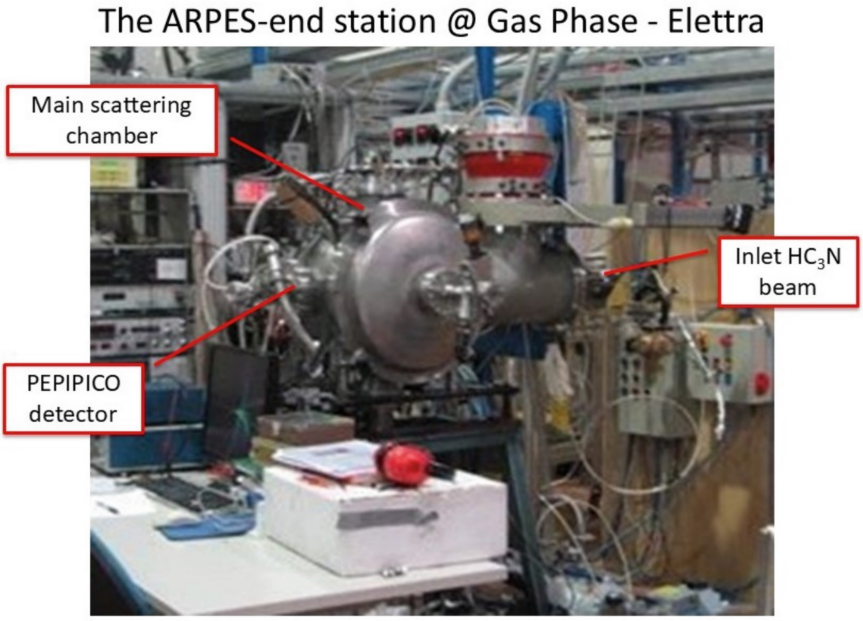


Fig. 2. The Angle-Resolved PhotoEmission Spectroscopy (ARPES) experimental station which is available at the GasPhase beamline of the Elettra synchrotron facility (Trieste, Italy) in order to perform double photoionization studies on precursor molecules of environmental and astrochemical interest.

These experimental studies have relied primarily on the use of our PEPIPICO detector coupled to a TOF mass spectrometer. This experimental setup is an ion extraction and detection system that uses a position-sensitive detector. It consists of a stack of three microchannel plates with a multi-anode array arranged in 32 rows and 32 columns. This array was specifically designed to measure the spatial momentum components of fragment ions produced in double photoionization experiments, as our research group has done in the past, starting with the in-depth and comprehensive characterization of simple halogen-bonded molecules [19–21].

In Fig. 3 are reported: (a) the scheme of the GasPhase beamline, where is operating a Monochromator 400 lines/mm spherical grating in first order diffraction with entrance and exit slits adjusted in order to give a photon energy resolution of 1.5 meV; (b) a picture

of the PEPIPICO device used in our experimental determinations whose characteristics are fully described elsewhere [22].

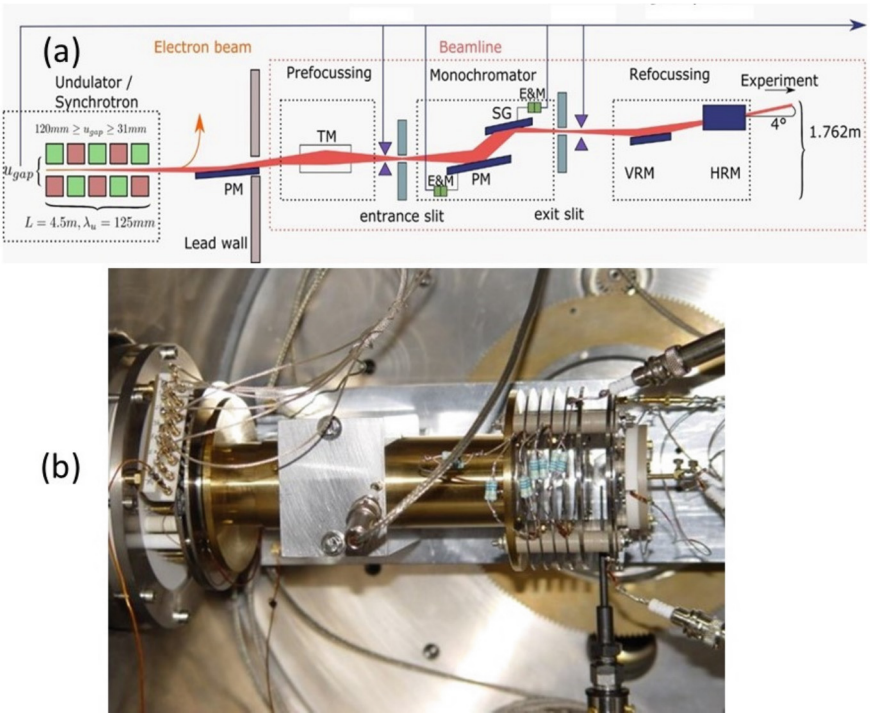


Fig. 3. (a) The scheme of the GasPhase beamline, reproduced courtesy of the Elettra Synchrotron Facility (Italy). (b) A photo of the photoelectron-photoion-photoion coincidence (PEPIPICO) device used in our experiments.

For the GasPhase beamline experiment it was necessary to use 6 shifts (48 h) of beam time in which the cyanoacetylene sample, after being produced using the above-mentioned synthesis procedure and schematized in Fig. 1, was easily employed to produce an effusive molecular beam, being safely stored at −20 °C at the Elettra chemistry laboratory.

3 Computational

The Fig. 4 shows the coincidence spectrum obtained in the double photoionization of cyanoacetylene by PEPIPICO, recorded at a photon energy of 45.0 eV. The figure, in red, also reports the mass spectra of the ions formed in both the single and double photoionization events. This kind of spectrum is typical of any coincidence plot recorded in double photoionization experiments with the PEPIPICO technique. In particular, t_1 and t_2 are the two time-of-flight values for any pair of fragment ions produced by Coulomb explosion of the short-lived HC_3N^{2+} molecular dication: they are related to

the blue dots in the diagram because they are generated by the same photoionization event. Indeed, all dots in the plot of coincidence of Fig. 4 are related to the coincidence events as a function of the arrival times, t_1 and t_2, of the two fragments ions to the detector of our TOF mass spectrometer. It has to be noted that the diagonal traces in the diagram are generated from "false" and spurious coincidences that must be neglected in the data analysis which is conducted using the Monte Carlo computational model specially developed in our laboratory (see below).

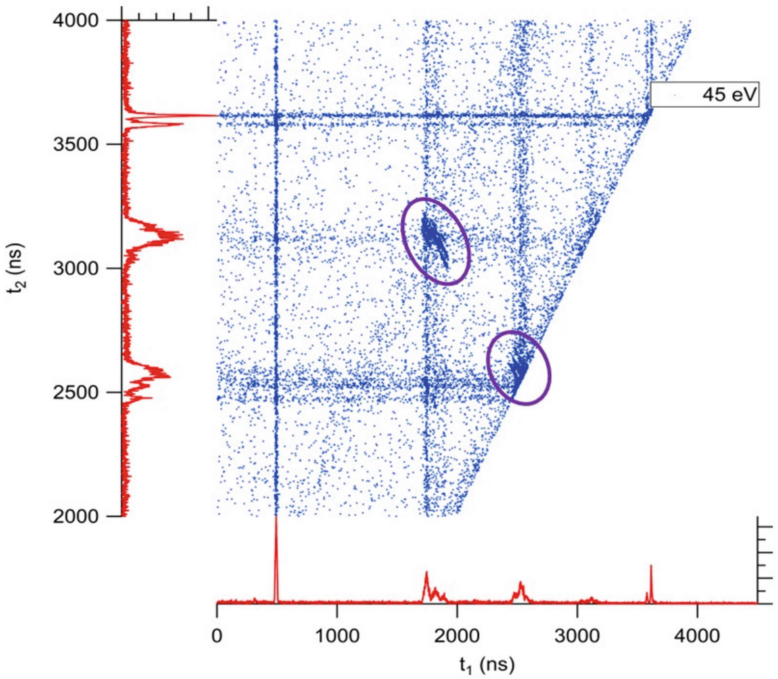


Fig. 4. The coincidence plot recorded in the double photoionization of HC_3N at a photon energy of 45.0 eV. The coincidence signals of the major open fragmentation channels are highlighted in the figure as the typical islands inside the purple ovals (see also Eqs. (2)–(10) and Sect. 4). In the figure t_1 and t_2 are respectively the arrival times of the first and second fragment ions at the detector.

In general, the analysis of PEPIPICO spectra (such as the diagram in Fig. 4), recorded as a function of synchrotron radiation photon energy, allows us to obtain important observables that are essential to fully characterize the energetics and dynamics of fragmentation that occurs following the double photoionization of the neutral precursor HC_3N. These observables are: i) the determination of all possible fragmentation channels with their corresponding threshold energies; ii) the measurement of their relative cross sections as a function of light energy; iii) the determination of the kinetic energy released (KER) of the two product ion fragments formed in each open dissociation channel; iv) the angular distribution of the product ions with respect to the direction of the

polarization vector of light; v) the evaluation of the lifetime of the intermediate dication that leads to the Coulomb explosion and subsequent fragmentation processes. This computational procedure can be performed by analyzing the shape and size of the peaks and islands (see those highlighted inside the purple ovals in Fig. 4) recorded for each ion pair produced in the coincidence spectrum, as proposed by J.H.D. Eland's group [25] who has been a leader in multiple photoionization studies.

As mentioned above, in order to confirm the reliability of our speculations on the fragmentation dynamics of the investigated systems, in our laboratory we generally perform an additional computational test, based on an alternative Monte Carlo trajectory simulation. Our computational routine is based on the determination of: i) the distribution of the delay times t_1 and t_2 of the fragment ions arriving at the TOF detector; ii) the number of recorded coincidences related to the formation of the metastable molecular dication as a function of the difference (t_2-t_1). Subsequently, a best fit of these data is performed by a Monte Carlo simulation of the ion trajectories (in general, using 10^6 trajectories) inside the TOF mass spectrometer. Finally, by minimizing the mean square values, χ^2, of the difference between the experimental data and the Monte Carlo simulation, it is possible to evaluate the lifetime of the metastable dication with its formation energy threshold. A similar procedure is applied to analyze both the distributions of the kinetic energy released (KER) of the fragment ions and their angular distributions as a function of the investigated photon energy. The reader interested in all the details of this computational procedure can refer to specific articles already published [23, 24].

4 Results and Discussion

Preliminary results concerning the double photoionization of the cyanoacetylene HC_3N molecule by synchrotron radiation in the 30–50 eV photon energy range are here presented. During the assigned beamtime at the GasPhase beamline of the Elettra Synchrotron Facility (Trieste, Italy) it was possible to perform the experiment successfully fully exploring the fragmentation microscopic dynamics of the intermediate $(HC_3N)^{2+}$ dication. The Fig. 5 shows a typical mass spectrum recorded at 45.0 eV, where all fragmentation channels are open.

Exploiting the photoelectron-photoion-photoion coincidence (PEPIPICO) device mounted in the ARPES end station (see Sect. 2 and Fig. 2), it was possible to measure accurately the threshold energy for the formation of the cyanoacetylene $(HC_3N)^{2+}$ molecular dication, obtaining the value of 31.4 ± 0.1 eV, and for the various fragmentation channels produced by double photoionization of HC_3N (see Eqs. (2)–(10) below). The measured threshold energies were obtained by accumulating 1–2 h at each investigated photon energy with a 100–200 meV energy steps. All recorded fragmentation channels and their corresponding threshold photon energies are reported in Eqs. (2)–(10) below:

$$HC_3N + h\nu \rightarrow C^+ + HCCN^+ + 2e^- \qquad h\nu \geq 33.5 \pm 0.2\ \text{eV} \tag{2}$$

$$\rightarrow CH^+ + CCN^+ + 2e^- \qquad h\nu \geq 33.5 \pm 0.2\ \text{eV} \tag{3}$$

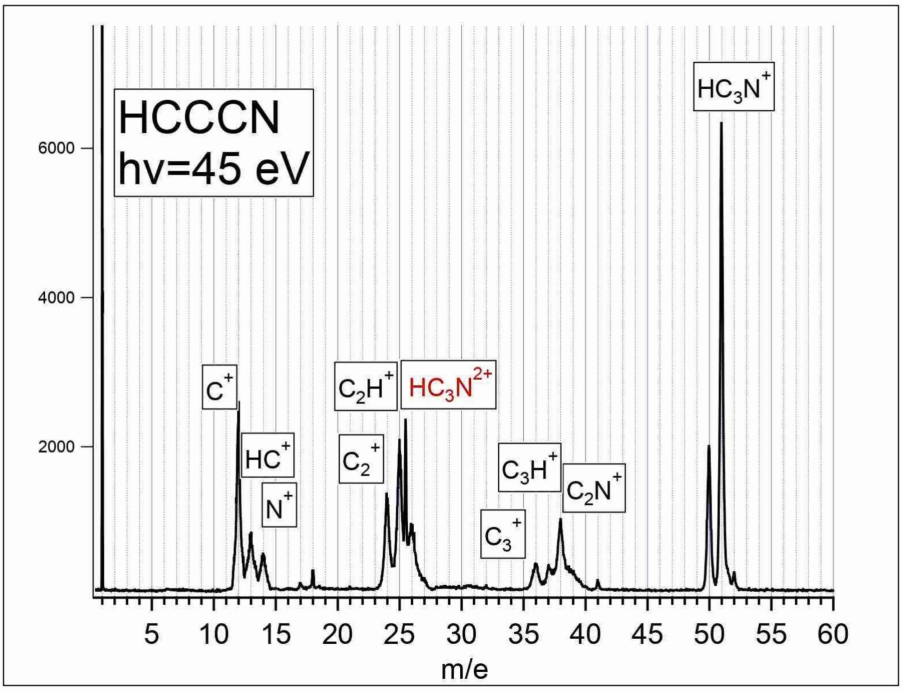


Fig. 5. The mass spectrum recorded in the double photoionization of HC_3N at a photon energy of 45.0 eV. In red is highlighted the m/e peak related to cyanoacetylene molecular dication. (Color figure online)

$$\rightarrow \mathrm{CCCN^+ + H^+ + 2e^-} \qquad h\nu \geq 35.8 \pm 0.2\ \mathrm{eV} \tag{4}$$

$$\rightarrow \mathrm{C^+ + CCN^+ + H + 2e^-} \qquad h\nu \geq 37.5 \pm 0.2\ \mathrm{eV} \tag{5}$$

$$\rightarrow \mathrm{CCCH^+ + N^+ + 2e^-} \qquad h\nu \geq 37.6 \pm 0.2\ \mathrm{eV} \tag{6}$$

$$\rightarrow \mathrm{CN^+ + CC^+ + H + 2e^-} \qquad h\nu \geq 37.8 \pm 0.2\ \mathrm{eV} \tag{7}$$

$$\rightarrow \mathrm{CCH^+ + CN^+ + 2e^-} \qquad h\nu \geq 38.0 \pm 0.2\ \mathrm{eV} \tag{8}$$

$$\rightarrow \mathrm{CCN^+ + H^+ + C + 2e^-} \qquad h\nu \geq 38.2 \pm 0.2\ \mathrm{eV} \tag{9}$$

$$\rightarrow \mathrm{CCC^+ + N^+ + H + 2e^-} \qquad h\nu \geq 38.5 \pm 0.2\ \mathrm{eV} \tag{10}$$

The measured threshold energies reported above were obtained by accumulating 6–8 h at each investigated photon energy with a 200 meV photon energy step. They were determined by using the Wannier procedure as done in previous papers [26–29].

Furthermore, by analyzing recorded coincidence plots at each investigated photon energy (see Fig. 2 – lower panel) by using the procedure described in Sects. 2 and 3, we should be able to extract the kinetic energy released (KER) for each fragment ion formation reported above as well as the relative cross sections of all the open channels (see Eqs. (2)–(10)). For this purpose, a careful analysis of the collected data is needed and this work will be done as soon as possible in order to fully characterize the microscopic dynamics of all recorded fragmentation reactions (see Eqs. (2)–(10) above).

5 Conclusions and Future Remarks

Some results obtained in a recent beamtime performed at the GasPhase beamline of the Elettra Synchrotron Facility (Trieste, Italy) and related to the double photoionization of the cyanoacetylene HC_3N molecule studied in the 30–50 eV photon energy range have been presented. A preliminary analysis based on a Monte Carlo computational routine was performed which allowed to determine the formation energy threshold for the intermediate molecular dication cyanoacetylene $(HC_3N)^{2+}$: a value of 31.4 ± 0.1 eV was measured. Moreover, such a molecular dication once formed is able to give rise to Coulomb explosion with subsequent formation of nine fragmentation channels. They were studied using photoelectron-photoion-photoion coincidence spectroscopy (PEPIPICO) coupled to time-of-flight (TOF) mass spectrometry, and the energy threshold was measured for each. Furthermore, in our ongoing work interpreting the experimental data, through a more detailed analysis of the coincidence plots recorded at each photon energy studied, we should be able to extract: i) the relative cross sections of the observed fragmentation channels as a function of photon energy; ii) the kinetic energy released for the formation of each fragment ion.

In conclusion, the new information obtained and that expected from a comprehensive and detailed analysis of the experimental data should allow us to comprehensively characterize the energetics and dynamics of the dissociation of cyanoacetylene dications due to the direct double photoionization of the precursor molecules. This is a process of astrochemical importance, as it can be induced by the interaction of neutral cyanoacetylene molecules with cosmic rays, UV and EUV photons, and synchrotron radiation, all of which are ionizing energy sources present in space. The direct approach used in this work represents the most effective method for studying the energetics of the $(HC_3N)^{2+}$ dication and the topology of its potential energy surfaces (PES). This, as already done for other systems by our research group [30–35], will constitute a further objective to be achieved. Finally, the study proposed here constitutes a first valid attempt, which will be further explored, to investigate the role of excited electronic states on the reactivity of the cyanoacetylene dication, a relevant species in space prebiotic chemistry.

Our research group, which has been involved since 2001 at the GasPhase and CiPo beamlines of the Elettra synchrotron radiation facility (see Fig. 6) in experiments studying elementary chemical processes of atmospheric and astrochemical interest [36, 37], will continue to contribute to the future development of the new MOST beamline currently under construction. We will do so by carrying on the legacy left by our colleague, collaborator, great scientist and friend Robert Richter, who has made a fundamental contribution to the development of scientific research carried out at Elettra over the past thirty years leaving an indelible mark.

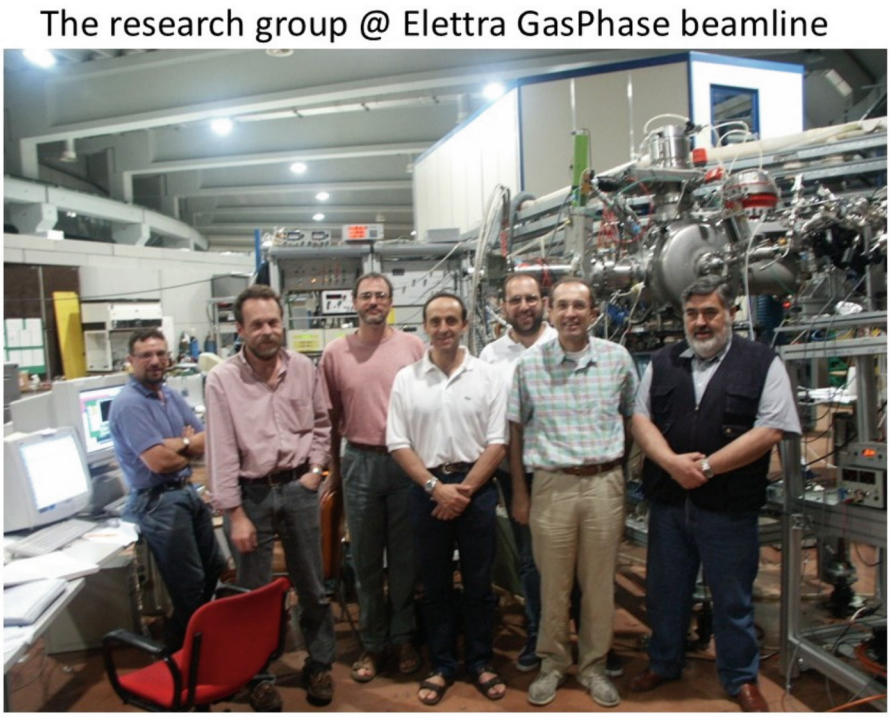


Fig. 6. From left to right: Michele Alagia, Robert Richter, Stefano Falcinelli, Stefano Stranges, Pietro Candori, Brunetto Brunetti and Franco Vecchiocattivi, during a beamtime at the GasPhase beamline of Elettra (Trieste) in the summer of 2001.

6 Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments. Dr. Robert Richter, a valued friend and collaborator, died after carrying out the experiments described, giving his ever-fundamental contribution. This work and our future efforts are dedicated to him, whose intellectual honesty and genuine love for physics and chemistry continue to inspire our scientific life. This work has been funded by the European Union – NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem (grant number ECS00000041) - VITALITY - CUP J97G22000170005. M.P. acknowledges the Italian Space Agency (ASI, DC-VUM-2017-034, Grant no. 2019-3 U.0 Life in Space).

References

1. Sanchez, R.A., Ferris, J.P., Orgel, L.E.: Cyanoacetylene in prebiotic synthesis. Science **154**(3750), 784–785 (1966)

2. Robertson, M.P., Miller, S.L.: An efficient prebiotic synthesis of cytosine and uracil. Nature **375**, 772–774 (1995)
3. Thissen, R., Witasse, O., Dutuit, O., et al.: Doubly-charged ions in the planetary ionospheres: a review. Phys. Chem. Chem. Phys. **13**(41), 18264–18287 (2011)
4. Brunetti, B., Candori, P., Falcinelli, S., Pirani, F., Vecchiocattivi, F.: The stereodynamics of the penning ionization of water by metastable neon atoms. J. Chem. Phys. **139**(16), 164305 (2013)
5. Falcinelli, S., Rosi, M., Candori, P., Vecchiocattivi, F., et al.: Modeling the intermolecular interactions and characterization of the dynamics of collisional autoionization processes. In: Murgante, B., et al. (eds.) ICCSA 2013. LNCS, vol. 7971, pp. 69–83. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39637-3_6
6. Falcinelli, S., et al.: Molecular dications in planetary atmospheric escape. Atmosphere **7**, 112 (2016)
7. Falcinelli, S., Rosi, M.: Production and characterization of molecular dications: experimental and theoretical efforts. Molecules **25**, 4157 (2020)
8. Sabzyan, H., Keshavarz, E., Noorisafa, Z.: Diatomic dications and anions. J. Iran. Chem. Soc. **11**, 871 (2014)
9. Gronoff, G., Lilensten, J., Simon, C., Witasse, O., et al.: Modelling dications in the diurnal ionosphere of Venus. Astron. Astrophys. **465**, 641–645 (2007)
10. Falcinelli, S., Pirani, F., Vecchiocattivi, F.: The possible role of penning ionization processes in planetary atmospheres. Atmosphere **6**(3), 299–317 (2015)
11. Lilensten, J., et al.: Prediction of a N_2^{++} layer in the upper atmosphere of Titan. Geophys. Res. Lett. **32**, 03203 (2005)
12. Salén, P., Schio, L., Richter, R., Alagia, M., Stranges, et al.: Electronic state influence on selective bond breaking of core-excited nitrosyl chloride (ClNO). J. Chem. Phys. **157**(12), 124306 (2022)
13. Witasse, O., Dutuit, O., Lilensten, J., et al.: Prediction of a CO_2^{2+} layer in the atmosphere of Mars. Geophys. Res. Lett. **30**, 1360 (2003)
14. Falcinelli, S., Rosi, M., Candori, P., et al.: Kinetic energy release in molecular dications fragmentation after VUV and EUV ionization and escape from planetary atmospheres. Planet. Space Sci. **99**, 149–157 (2014)
15. Falcinelli, S., Pirani, F., Alagia, M., et al.: The escape of O^+ ions from the atmosphere: an explanation of the observed ion density profiles on Mars. Chem. Phys. Lett. **666**, 1–6 (2016)
16. Miller, F.A., Lemmon, D.H.: The infrared and Raman spectra of dicyanodiacetylene. Spectrochim. Acta **23A**, 1415 (1967)
17. Parriani, M., et al.: Synthesis of chemical species relevant in environmental and astro chemistry: their production and characterization by synchrotron radiation. In: Gervasi, O., et al. (eds.) ICCSA 2025. LNCS, vol. 15888, pp. 300–314. Springer, Heidelberg (2026). https://doi.org/10.1007/978-3-031-97596-7_20
18. Teixidor, M.M., Pirani, F., Candori, P., Falcinelli, S., Vecchiocattivi, F.: Predicted structure and energetics of HCl^{2+}. Chem. Phys. Lett. **379**(1–2), 139–146 (2003)
19. Alagia, M., Brunetti, B.G., Candori, P., et al.: Threshold-photoelectron-spectroscopy-coincidence study of the double photoionization of HBr. J. Chem. Phys. **120**(15), 6980–6984 (2004)
20. Bartocci, A., Belpassi, L., Cappelletti, D., et al.: Catching the role of anisotropic electronic distribution and charge transfer in halogen bonded complexes of noble gases. J. Chem. Phys. **142**(18), 184304 (2015)
21. Pirani, F., Cappelletti, D., Falcinelli, S., et al.: Selective emergence of halogen bond in ground and excited states of noble-gas—chlorine systems. Angew. Chem. **58**(13), 4195–4199 (2019)
22. Lavollée, M.: A new detector for measuring three-dimensional momenta of charged particles in coincidence. Rev. Sci. Instrum. **70**(7), 2968–2974 (1999)

23. Falcinelli, S., Vecchiocattivi, F., Pirani, F., et al.: The fragmentation dynamics of simple organic molecules of astrochemical interest interacting with VUV photons. ACS Earth Space Chem. **3**, 1862–1872 (2019)
24. Parriani, M., Giustini, A., Alagia, M., et al.: Data analysis of molecular dications fragmentation processes by double photoionization with synchrotron radiation. In: Gervasi, O., et al. (eds.) ICCSA 2024. LNCS, vol. 14818, pp. 55–70. Springer, Heidelberg (2024)
25. Field, T.A., Eland, J.H.D.: Lifetimes of metastable molecular doubly charged ions. Chem. Phys. Lett. **211**, 436–442 (1993)
26. Wannier, G.H.: The threshold law for single ionization of atoms or ions by electrons. Phys. Rev. **90**(5), 817–825 (1953)
27. Ben Arfa, M., Lescop, B., Cherid, M., Brunetti, B., et al.: Ionization of ammonia molecules by collision with metastable neon atoms. Chem. Phys. Lett. **308**(1–2), 71–77 (1999)
28. Biondini, F., Brunetti, B.G., Candori, P., De Angelis, F., et al.: Penning ionization of N_2O molecules by $He^*(2^{3,1}S)$ and $Ne^*(^3P_{2,0})$ metastable atoms: theoretical considerations about the intermolecular interactions. J. Chem. Phys. **122**(16), 164308 (2005)
29. Falcinelli, S., Pirani, F., Vecchiocattivi, F.: Adiabatic and nonadiabatic effects in the transition states of state to state autoionization processes. Phys. Rev. Lett. **121**(16), 163403 (2018)
30. Brunetti, B., Candori, P., Falcinelli, S., Lescop, B., et al.: Energy dependence of the Penning ionization electron spectrum of $Ne^*(^3P_{2,0})$+Kr. Eur. Phys. J. D **38**(1), 21–27 (2006)
31. Alagia, M., Brunetti, B.G., Candori, P., et al.: Low-lying electronic states of HBr^2+. J. Chem. Phys. **120**(15), 6985–6991 (2004)
32. Falcinelli, S., Farrar, J.M., Vecchiocattivi, F., Pirani, F.: Quantum-state controlled reaction channels in chemi-ionization processes: radiative (optical−physical) and exchange (oxidative−chemical) mechanisms. Acc. Chem. Res. **53**(10), 2248–2260 (2020)
33. Brunetti, B., Candori, P., De Andres, J., et al.: Dissociative ionization of methyl chloride and methyl bromide by collision with metastable neon atoms. J. Phys. Chem. A **101**(41), 7505–7512 (1997)
34. Alagia, M., Biondini, F., Brunetti, B.G., Candori, P., et al.: The double photoionization of HCI: an ion-electron coincidence study. J. Chem. Phys. **121**(21), 10508–10512 (2004)
35. Balucani, N., et al.: Collisional autoionization dynamics of $Ne^*(^3P_{2,0})$–H_2O. Chem. Phys. Lett. **546**, 34–39 (2012)
36. Alagia, M., et al.: Mass spectrometric study of double photoionization of HBr molecules. J. Chem. Phys. **117**(3), 1098–1102 (2002)
37. Alagia, M., et al.: The double photoionization of hydrogen iodide molecules. J. Chem. Phys. **124**(20), 204318 (2006)

Reactions of N(^{2}D) with Polycyclic Aromatic Hydrocarbons: A Computational Approach

Marzio Rosi[1,2(✉)], Nadia Balucani[3], Luca Mancini[3], Giacomo Pannacci[3], Marco Parriani[3], Dimitrios Skouteris[4], and Gianmarco Vanuzzo[3]

[1] Dipartimento di Ingegneria Civile ed Ambientale, Università di Perugia, 06125 Perugia, Italy
marzio.rosi@unipg.it
[2] Istituto CNR di Scienze e Tecnologie Chimiche "Giulio Natta" (CNR-SCITEC), 06123 Perugia, Italy
[3] Dipartimento di Chimica, Biologia e Biotecnologie, Università di Perugia, 06123 Perugia, Italy
{nadia.balucani,luca.mancini,giacomo.pannacci,marco.parriani, gianmarco.vanuzzo}@unipg.it
[4] Master-Tec srl, 06128 Perugia, Italy

Abstract. Investigating large polycyclic aromatic hydrocarbons (PAHs) in interaction with atomic nitrogen poses a major computational challenge due to the high demands on processing time and temporary storage space. Highly precise approaches, such as CCSD(T) paired with extensive basis sets, are often impractical. This paper explores a cost-effective computational approach that delivers sufficiently reliable outcomes without excessive resource needs.

Specifically, we employ the ONIOM method developed by Morokuma and colleagues. This involves a two-layer ONIOM scheme, where the high-level layer applies the CCSD(T) method with the aug-cc-pVTZ basis set, while the low-level layer uses the hybrid B3LYP functional paired with the aug-cc-pVTZ basis set. The strategy is implemented to model the interaction between N(^{2}D) and both naphthalene and phenanthrene.

Keywords: Computational chemistry · ONIOM · Density Functional Theory methods · Coupled Cluster calculations · Potential energy surface

1 Introduction

Atomic nitrogen in its ground state (^{4}S) exhibits very low reactivity toward closed-shell molecules such as hydrocarbons, whereas in its first electronically excited ^{2}D state, it displays substantial reactivity with hydrocarbons. N(^{2}D) was detected in the water-poor comet C/2016 R2 (Pan-STARRS) [1] and in a plethora of strongly photon-irradiated environments including the Orion Nebula (M42), low-ionization H II regions (M43), planetary nebulae (*i.e.* the Ring Nebula), supernova remnants (*i.e.* the Crab Nebula), and Herbig-Haro objects [2–5]. Polycyclic aromatic hydrocarbons and similar compounds are thought to be ubiquitous throughout the interstellar medium (ISM), with aromatic chemistry playing a prominent role from the initial phases of star formation.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 448–460, 2027.
https://doi.org/10.1007/978-3-032-30521-3_28

Nitrogen, in its molecular form, and hydrocarbons, both aliphatic and aromatic, are also the main components of the atmosphere of Titan [6, 7]. This atmosphere is similar, in some aspects, to the primordial atmosphere of Earth [8, 9] and for this reason has been extensively studied by several missions [10, 11]. Among the hydrocarbons identified on Titan there is benzene [12, 14]. Dinitrogen in the atmosphere of Titan can dissociate into atomic nitrogen both in its ground state or ^{2}D excited state in similar amounts [15, 16] and N (^{2}D) can easily react with other constituents of the upper atmosphere of Titan or with species present in the ISM [17–24].

The reaction of N(^{2}D) with several aliphatic hydrocarbons of increasing complexity, starting from methane, has been already investigated both theoretically, using accurate ab initio methods and kinetic calculations, and experimentally, through crossed molecular beams techniques [25–32]. A similar approach is employed for the reactions between N(^{2}D) and aromatic hydrocarbons, starting from benzene [33–35] and toluene [36, 37] with the intention to extend the study to more complex aromatic systems, like polycyclic aromatic hydrocarbons (PAHs).

In this contribution we will present a computational strategy to study large systems as those relative to the interaction of nitrogen in its excited ^{2}D state and polycyclic aromatic hydrocarbons, in particular naphthalene and phenanthrene. We will use the previously investigated system N(^{2}D) + C_6H_6 as a benchmark calculation in order to establish the better approach for the investigation of larger systems [35, 38].

2 Computational Details

The optimization of the stationary points of the Potential Energy Surface (PES) of N(^{2}D) interacting with benzene, naphthalene and phenanthrene has been performed at Density Functional Theory (DFT) level of calculation using the B3LYP [39, 40] functional and the 6–311 + G(d,p) basis set [41, 42], with the addition of the D3 dispersion [43]. For benzene the DFT geometry optimizations were performed also using the aug-cc-pVTZ basis set [44]. In order to assign the saddle points to the relative reactants and products, we performed intrinsic reaction coordinate (IRC) calculations [45, 46]. The energy of all the stationary points of N(^{2}D) + benzene and naphthalene was computed also at the CCSD(T) level [47–49] using the correlation consistent aug-cc-pVTZ basis set [44]. The energy of N(^{2}D) has been estimated by adding the experimental [50] separation N(^{4}S) – N(^{2}D) of 230 kJ mol^{-1} to the energy of N(^{4}S) at all levels of calculations. The zero point energy correction (ZPE) computed using the DFT vibrational frequencies has been added at all level of calculation. A similar computational scheme has been previously successfully applied [51–56].

More precise computations were carried out for the stationary points of the N(^{2}D) + benzene system at the CCSD(T) level. These included corrections from density-fitted (DF) MP2 extrapolation to the complete basis set, along with corrections for core electron excitations. In particular, we compute the energies as:

$$E_{CCSD(T)/CBS} = E(CCSD(T)/aug-cc-pVTZ) + \left[E(CCSD(T,core)/cc-pVTZ) - E(CCSD(T)/cc-pVTZ)\right] + \left[E(DF-MP2/CBS) - E(DF-MP2/aug-cc-pVTZ)\right]$$

where E(DF-MP2/CBS) is defined as:

$$\mathrm{E(DF-MP2/CBS)} = \mathrm{E}\big[\,(\mathrm{DF-MP2/aug-cc-pVQZ})\big] + 0.5772* \\ [\mathrm{E(DF-MP2/aug-cc-pVQZ)} - \mathrm{E(DF-MP2/aug-cc-pVTZ)}]$$

The E(DF-MP2/CBS) extrapolation was performed using Martin's two parameter scheme [57]. These energies will be denoted as CCSD(T)/CBS although they include also the core-valence correlation correction. The accuracy of these calculations should be within ± 5 kJ/mol. This approach is computationally unfeasible for the systems $N(^2D)$ + naphthalene or phenanthrene. For this last system also CCSD(T)/aug-cc-pVTZ calculations are unfeasible, being too expensive as far as computational resources are concerned.

For these reasons, we decided to explore the opportunities provided by the ONIOM approach for the investigation of $N(^2D)$ + naphthalene and phenanthrene. We used the ONIOM method of Morokuma and co-workers [58–60]. We used a two level ONIOM approach, with the hybrid B3LYP approach in conjunction with the aug-cc-pVTZ basis set being used for the low level treatment and the CCSD(T)/aug-cc-pVTZ for the high level treatment. The high level treatment is applied to nitrogen and the ring involved in the interaction, while the rest of the PAH is described at the low level treatment. The dangling bonds are replaced by hydrogens. The computational cost of this treatment is comparable to that of the system $N(^2D) + C_6H_6$.

The DFT and CCSD(T) calculations were done with Gaussian 09 [61] and the CCSD(T)/CBS calculations with Molpro [62, 63], while the analysis of the vibrational frequencies was performed using Molekel [64, 65].

3 Results and Discussion

3.1 $N(^2D)$ + Benzene

In Fig. 1 we have reported the Potential Energy Surface (PES) relative to the lowest exit channel of the reaction of $N(^2D)$ and benzene. The complete PES has been previously published [35]. The approach of $N(^2D)$ to the benzene ring gives rise to a first intermediate (B-w1) which evolves through a low transition state to a more stable species (B-w2) with the nitrogen bonded to two carbon atoms. This species evolves with the insertion of nitrogen into the C—C bond giving rise to a 7-membered ring (B-w3). The 7-membered ring evolves through minima B-w4 and B-w5 giving rise to the final products, HCN and cyclopentadienyl radical. In Fig. 1 we have reported the energies relative to $N(^2D) + C_6H_6$ computed at different level of accuracy: B3LYP with an aug-cc-pVTZ basis set [44], coupled cluster with inclusion of triple excitations (CCSD(T)) with the same aug-cc-pVTZ basis set and CCSD(T) with an extrapolation to complete basis set limit (CBS) and inclusion also of the core-valence correlation effects. This last approach is extremely accurate and provides results with an error which can be estimated lower than ± 5 kJ mol^{-1}. However, this accuracy has a very high computational cost in terms of CPU time and disk space. For this reason this approach can be applied only to relatively small systems and cannot be applied, for instance, to the study of the interaction of $N(^2D)$ with polycyclic aromatic hydrocarbons. We can use, however, these results as benchmark calculations in order to choose the best approach to the study of larger systems.

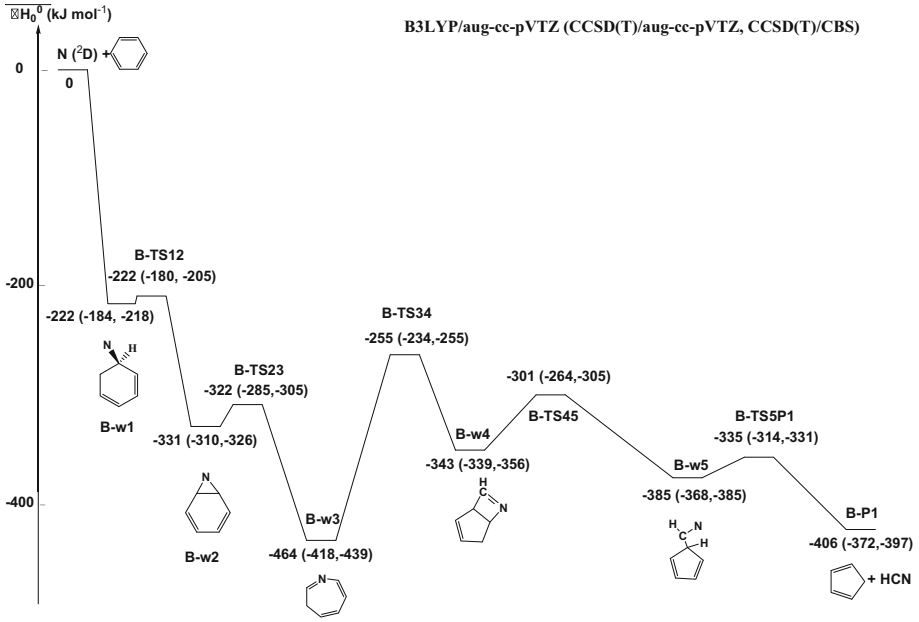


Fig. 1. PES for the N(^{2}D) + benzene reaction. Relative energies (kJ mol^{-1}) with respect to N(^{2}D) + benzene, computed at B3LYP/aug-cc-pVTZ, CCSD(T)/aug-cc-pVTZ and CCSD(T)/CBS level. Zero point energy correction always included.

From the values reported in Fig. 1 we can notice that CCSD(T)/aug-cc-pVTZ relative energies are higher than B3LYP energies by an average of 28 kJ mol^{-1}, while CCSD(T)/CBS energies are lower than CCSD(T) energies by an average of 23 kJ mol^{-1}. Indeed, most of the times there is a better agreement between B3LYP and CCSD(T)/CBS results with respect to CCSD(T)/aug-cc-pVTZ and CCSD(T)/CBS results. Unfortunately, the discrepancy between B3LYP and more accurate results is not always the same. For this reason it is necessary to perform more accurate calculations as CCSD(T), but we must be aware that CCSD(T) calculations converge on basis set very slowly and CCSD(T) calculations even with an extended basis set as the aug-cc-pVTZ show a basis set incompleteness error. In this case this error is around 20 kJ mol^{-1}.

3.2 N(^{2}D) + Naphthalene

In Fig. 2 we have reported the geometries of the stationary points of the lowest exit channel of the PES relative to the reaction of N(^{2}D) with naphthalene, optimized at B3LYP/6–311 + G(d,p) level. The PES relative to the lowest exit channel has been reported in Fig. 3, while the complete PES has been published elsewhere [66].

The interaction of N(^{2}D) with naphthalene gives rise to species N-w3 where the nitrogen atom is already inserted into the C—C bond with the formation of a 7-membered ring. This species evolves through minima N-w4 and N-w5 to the exit channel where we have the formation of HCN. This reaction, except for the initial approach of nitrogen to the naphthalene, is very similar to the reaction of nitrogen and benzene. The second ring of

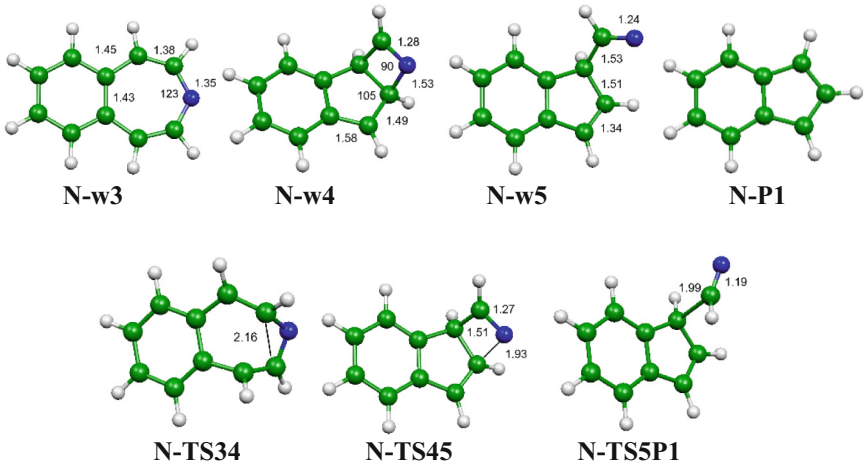


Fig. 2. Optimized geometries of the stationary points for the N(^{2}D) + naphthalene PES, computed at B3LYP/6-311 + G(d,p) level. Bond lengths in Å, angles in °.

naphthalene indeed acts as a spectator. This is confirmed by the optimized geometries of naphthalene and N-w3 reported in Fig. 4. From Fig. 4 we can notice that the geometrical parameters of the second ring (that is the ring not interacting with nitrogen) are almost the same in naphthalene and in the N-w3 minimum where the 7-membered ring is already formed. This consideration suggests a new computational approach for this system. We decided therefore to use a two level computational method as ONIOM, describing at an higher level as CCSD(T)/aug-cc-pVTZ the reactive part of the system (in this case the ring interacting with nitrogen and the nitrogen itself) and at a lower level as B3LYP/aug-cc-pVTZ the spectator ring. The computational cost in this case is comparable to that afforded for the system N(^{2}D) + benzene. In Fig. 3 we have reported the ONIOM energies in parentheses. For comparison we have performed also CCSD(T)/aug-cc-pVTZ calculations for the whole system and we have obtained energy values very close to those obtained at the ONIOM level. This confirms that the results of the ONIOM calculations are comparable to the results obtained applying the higher level method to the whole system. The results obtained with ONIOM of course present an error due to basis set incompleteness, but we can estimate this error using the N(^{2}D) + benzene benchmark calculations.

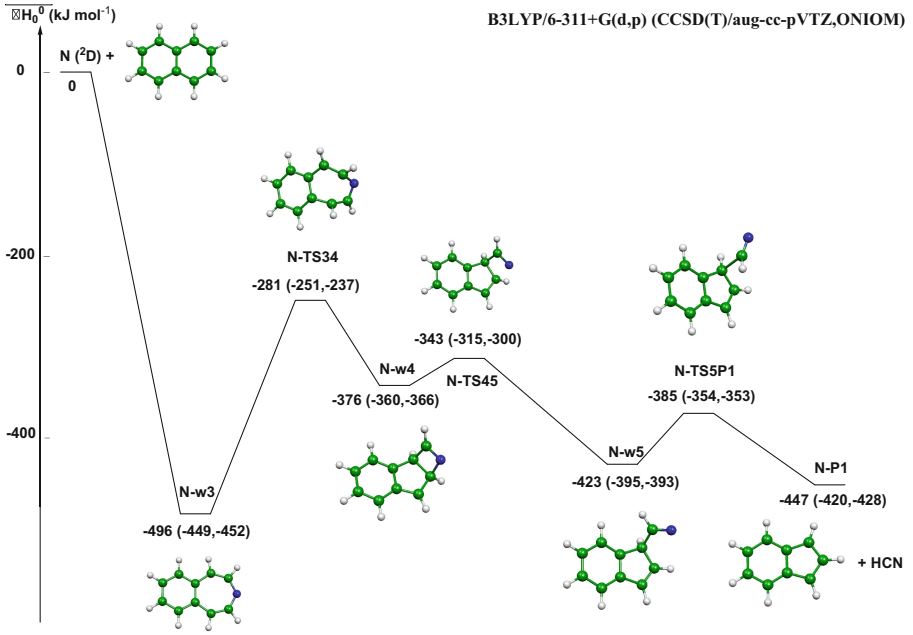
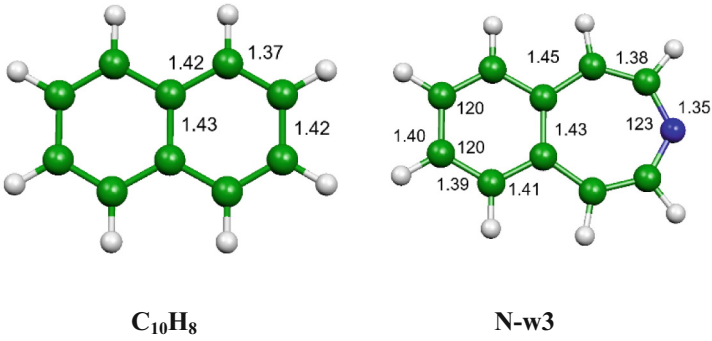


Fig. 3. PES for the N(^{2}D) + naphthalene reaction. Relative energies (kJ mol^{-1}) with respect to N(^{2}D) + naphthalene, computed at B3LYP/6–311 + G(d,p), CCSD(T)/aug-cc-pVTZ and ONIOM level. Zero point energy correction computed at DFT level always included.



$C_{10}H_8$ **N-w3**

Fig. 4. Comparison between the optimized geometries of naphthalene and N-w3, that is the minimum formed from the interaction of N(^{2}D) with naphthalene. Bond lengths in Å, angles in °.

3.3 N(^{2}D) + Phenanthrene

In Fig. 5 we have reported the geometries of the stationary points of the lowest exit channel of the PES relative to the reaction N(^{2}D) and phenanthrene, optimized at B3LYP/6–311 + G(d,p) level. The PES relative to the lowest exit channel has been reported in Fig. 6, while the complete PES has been published elsewhere [66].

The interaction of N(^{2}D) with phenanthrene gives rise to species P-w2 where the nitrogen atom interacts with two carbon atoms. This species evolves to minimum P-w3 where the nitrogen atom is inserted into the C—C bond with the formation of a 7-membered ring. This species evolves through minima P-w4 and P-w5 to the exit channel where we have the formation of HCN. This reaction, except for the initial approach of nitrogen to phenanthrene, is very similar to the reaction of nitrogen and benzene. The lateral rings of phenanthrene indeed act as spectators. This is confirmed by the optimized geometries of phenanthrene and P-w2 reported in Fig. 7. From Fig. 7 we can notice that the geometrical parameters of the second and third rings (that is the rings not interacting with nitrogen) are almost the same in phenanthrene and in the P-w2 minimum. We decided therefore to use also in this case a two level computational method as ONIOM, describing at an higher level as CCSD(T)/aug-cc-pVTZ the reactive part of the system (in this case the ring interacting with nitrogen and the nitrogen itself) and at a lower level as B3LYP/aug-cc-pVTZ the spectator rings. The computational cost in this case is comparable to that afforded for the system N(^{2}D) + benzene. In Fig. 6 we have reported the ONIOM energies in parentheses. The results obtained with ONIOM of course present an error due to basis set incompleteness, but we can estimate this error using the N(^{2}D) + benzene benchmark calculations.

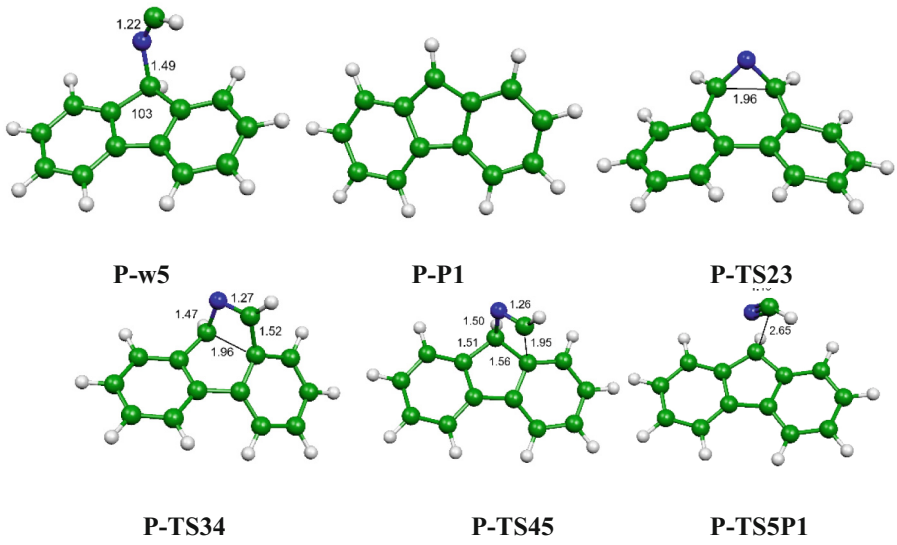


Fig. 5. Optimized geometries of the stationary points for the N(^{2}D) + phenanthrene PES, computed at B3LYP/6-311 + G(d,p) level. Bond lengths in Å, angles in °.

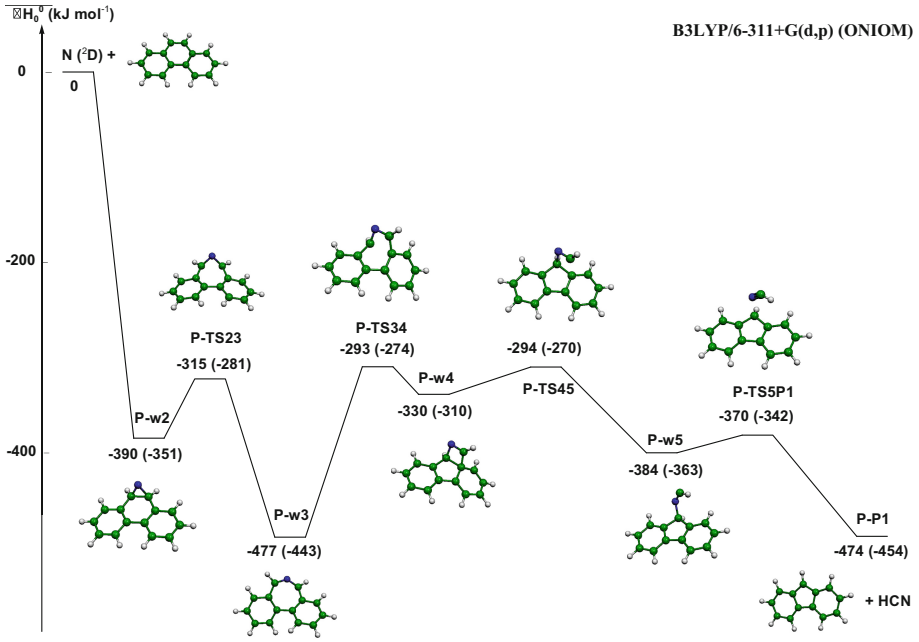
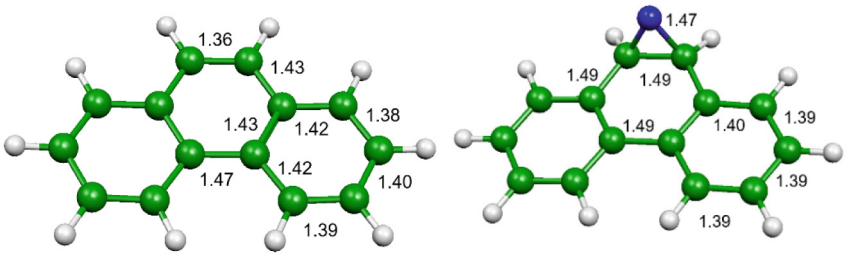


Fig. 6. PES for the N(^{2}D) + phenanthrene reaction. Relative energies (kJ mol^{-1}) with respect to N(^{2}D) + phenanthrene, computed at B3LYP/6-311 + G(d,p) and ONIOM level. Zero point energy correction computed at DFT level always included.



$C_{14}H_{10}$ **P-w2**

Fig. 7. Comparison between the optimized geometries of phenanthrene and P-w2, that is the minimum formed from the interaction of N(^{2}D) with phenanthrene. Bond lengths in Å.

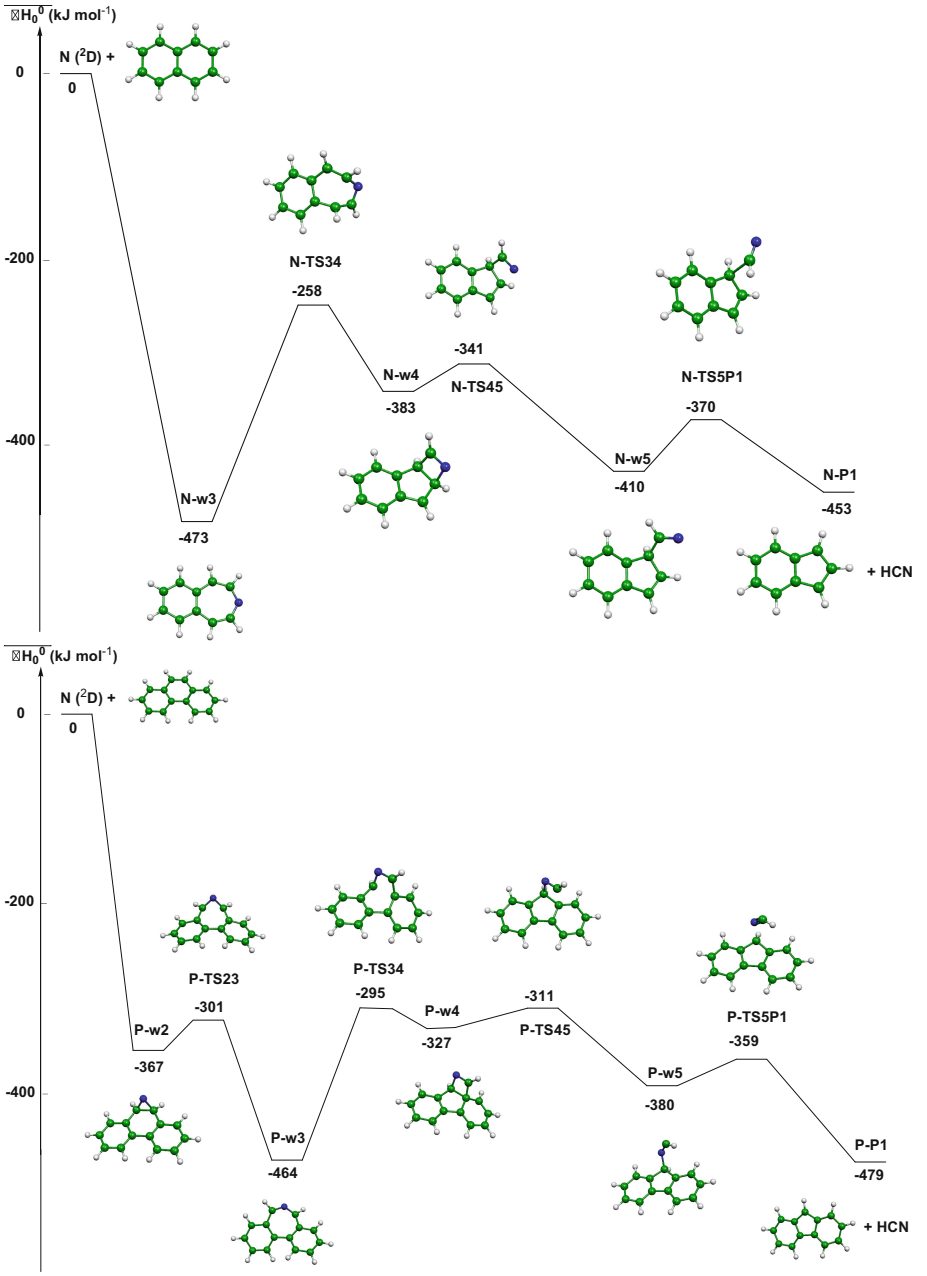


Fig. 8. Best estimate for the PES of N(^{2}D) + naphthalene (top) and phenanthrene (bottom). Relative energies (kJ mol^{-1}) with respect to N(^{2}D) + naphthalene or phenanthrene. Zero point energy correction computed at DFT level always included.

4 Conclusions

In Fig. 8 we have reported our best estimate for the lowest exit pathway for the reactions of N(^{2}D) with naphthalene (top) and phenanthrene (bottom). These values were obtained adding the difference between CCSD(T)/aug-cc-pVTZ and CCSD(T)/CBS energies computed for the system N(^{2}D) + benzene to the ONIOM energies computed for N(^{2}D) + naphthalene and N(^{2}D) + phenanthrene. Of course, we considered always stationary points (minima and transition states) of similar connectivity.

To sum up, the theoretical investigation—conducted at varying levels of complexity—of the interaction of N(^{2}D) with benzene, naphthalene, and phenanthrene points to a computational approach for studying large aromatic hydrocarbons interacting with nitrogen in its first electronic excited state ^{2}D. The use of the ONIOM approach with the description at an accurate level only of the reactive part of the system allows to extend the investigation also to complex PAHs. The use of benchmark calculations moreover allows to correct the final results in order to include also a correction for basis set incompleteness.

Acknowledgments. Support from the Italian Space Agency (Bando ASI Prot. N. DC-DSR-UVS-2022-231, Grant n.2023-10-U.0 MIGLIORA) is acknowledged. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 1409 published on 14.9.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title P20223H8CK Degradation of space technology polymers by thermospheric oxygen atoms and ions: an exploration of the reaction mechanisms at an atomistic level (ThermOPoly) – CUP J53D23014440001_- Grant Assignment Decree No. 1386 adopted on 01.09.2023 by the Italian Ministry of Ministry of University and Research (MUR). MR acknowledges financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title 2022JC2Y93 ChemicalOrigins: linking the fossil composition of the Solar System with the chemistry of protoplanetary disks – CUP J53D23001600006 - Grant Assignment Decree No. 962 adopted on 30.06.2023 by the Italian Ministry of Ministry of University and Research (MUR).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Raghuram, S., Hutsemekers, D., Opiton, C., Jehin, E., Bhadwaj, A., Manfroid, J.: Forbidden atomic carbon, nitrogen, and oxygen emission lines in the water-poor comet C/2016 R2 (Pan-STARRS). A&A **635**, A108 (2020)
2. Ferland, G.J., Henney, W.J., O'Dell, C.R., Porter, R.L., Van Hoof, P.A.M., Williams, R.J.R.: Pumping up the [N I] nebular lines. Astrophys. J. **757**, 79 (2012)
3. Dopita, M.A., Mason, D.J., Robb, W.D.: Atomic nitrogen as a probe of physical conditions in the interstellar medium. Astrophys. J. **207**, 102–109 (1976)
4. Ferland, G.J., Rees, M.J.: Radiative equilibrium of high-density clouds, with application to active galactic nucleus continua. Astrophys. J. **332**, 141 (1988)

5. Bautista, M.A.: Continuum fluorescence excitation of [N I] and [O I] lines in gaseous nebulae. Astrophys. J. **527**, 474–478 (1999)
6. Hörst, S. M.: Titan's atmosphere and climate. J. Geophys. Res. Planets **122**, 432482 (2017)
7. Vuitton, V., Yelle, R.V., Anicich, V.G.: The nitrogen chemistry of titan's upper atmosphere revealed. Astrophys. J. **647**, L175–L178 (2006)
8. Vuitton, V., et al.: Chemistry of Titan's atmosphere. In: Mueller-Wodarg, I., Griffith, C., Lellouch, E., Cravens, T. (eds.) Titan: Surface. Cambridge University Press, Atmosphere and Magnetosphere (2013)
9. Balucani, N.: Elementary reactions of N atoms with hydrocarbons: first steps towards the formation of prebiotic N-containing molecules in planetary atmospheres. Chem. Soc. Rev. **41**, 5473–5483 (2012)
10. Brown, R., Lebreton, J.P., Waite, J. (eds.): Titan from Cassini-Huygens. Springer, Netherlands (2010)
11. Lai, J.C.-Y., et al.: Mapping vinyl cyanide and other nitriles in Titan's atmosphere using ALMA. Astron. J. **154**(206), 1–10 (2017)
12. Vuitton, V., Yelle, R.V., Cui, J.: Formation and distribution of benzene on Titan. J. Geophys. Res. **113**, E05007 (2008)
13. Clark, R.N., et al.: Detection and mapping of hydrocarbon deposits on Titan. J. Geophys. Res. **115**, E10005 (2010)
14. Loison, J.C., Dobrijevic, M., Hickson, K.M.: The photochemical production of aromatics in the atmosphere of Titan. Icarus **329**, 55–71 (2019)
15. Lavvas, P., et al.: Energy deposition and primary chemical products in Titan's upper atmosphere. Icarus **213**, 233–251 (2011)
16. Dutuit, O., et al.: Critical review of N, N^+, N_2^+, N^{++} and N_2^{++} main production processes and reactions of relevance to Titan's atmosphere. Astrophys. J. Suppl. Ser. **204**, 20 (2013)
17. Balucani, N.: Nitrogen fixation by photochemistry in the atmosphere of titan and implications for prebiotic chemistry. In: Rodriguez, J.M., Raulin, F., Muller, C., Nixon, C. (eds.) The Early Evolution of the Atmospheres of Terrestrial Planets, Trigo- Springer Series in Astrophysics and Space Science Proceedings, vol. 35, p. 155 (2013)
18. Balucani, N.: Elementary reactions and their role in gas-phase prebiotic chemistry. Int. J. Mol. Sci. **10**, 2304–2335 (2009)
19. Imanaka, H., Smith, M.A.: Formation of nitrogenated organic aerosols in the Titan upper atmosphere. PNAS **107**, 12423–12428 (2010)
20. Balucani, N., et al.: Dynamics of the $N(^2D) + D_2$ reaction from crossed-beam and quasiclassical trajectory studies. J. Phys. Chem. A **105**, 2414–2422 (2001)
21. Balucani, N., et al.: Experimental and theoretical differential cross sections for the $N(^2D) + H_2$ reaction. J. Phys. Chem. A **110**, 817–829 (2006)
22. Homayoon, Z., Bowman, J.M., Balucani, N., Casavecchia, P.: Quasiclassical trajectory calculations of the $N(^2D) + H_2O$ reaction elucidating the formation mechanism of HNO and HON seen in molecular beam experiments. J. Phys. Chem. Lett. **5**, 3508–3513 (2014)
23. Balucani, N., Cartechini, L., Casavecchia, P., Homayoon, Z., Bowman, J.M.: A combined crossed molecular beam and quasiclassical trajectory study of the Titan-relevant $N(^2D) + D_2O$ reaction. Mol. Phys. **113**, 2296–2301 (2015)
24. Israel, G., et al.: Complex organic matter in Titan's atmospheric aerosols from in situ pyrolysis and analysis. Nature **438**, 796 (2005)
25. Balucani, N., et al.: Combined crossed molecular beam and theoretical studies of the $N(^2D) + CH_4$ reaction and implications for atmospheric models of Titan. J. Phys. Chem. A **113**, 11138–11152 (2009)
26. Balucani, N., et al.: Cyanomethylene formation from the reaction of excited nitrogen atoms with acetylene: a crossed beam and ab initio study. J. Am. Chem. Soc. **122**, 4443–4450 (2000)

27. Balucani, N., Cartechini, L., Alagia, M., Casavecchia, P., Volpi, G.G.: Observation of nitrogen-bearing organic molecules from reactions of nitrogen atoms with hydrocarbons: a crossed beam study of N(^{2}D) + ethylene. J. Phys. Chem. A **104**, 5655–5659 (2000)
28. Balucani, N., et al.: Formation of nitriles and imines in the atmosphere of Titan: combined crossed-beam and theoretical studies on the reaction dynamics of excited nitrogen atoms N(^{2}D) with ethane. Faraday Discuss. **147**, 189–216 (2010)
29. Balucani, N., et al.: Combined crossed beam and theoretical studies of the N(^{2}D) + C_2H_4 reaction and implications for atmospheric models of Titan. J. Phys. Chem. A **116**, 10467–10479 (2012)
30. Vanuzzo, G., et al.: Reaction N(^{2}D) + CH_2CCH_2 (Allene): an experimental and theoretical investigation and implications for the photochemical models of titan. ACS Earth Space Chem. **6**, 2305–2321 (2022)
31. Mancini, L., et al.: The reaction N(^{2}D) + CH_3CCH (methylacetylene): a combined crossed molecular beams and theoretical investigation and implications for the atmosphere of titan. J. Phys. Chem. A **125**, 8846–8859 (2021)
32. Mancini, L., de Aragao, E.V.F., Rosi, M., Skouteris, D., Balucani, N.: A theoretical investigation of the reactions of N(^{2}D) with small alkynes and implications for the prebiotic chemistry of titan. In: LNCS, vol. 12251, pp. 717–729 (2020)
33. Balucani, N., Pacifici, L., Skouteris, D., Caracciolo, A., Casavecchia, P., Rosi, M.: A theoretical investigation of the reaction N(^{2}D) + C_6H_6 and implications for the upper atmosphere of Titan. In: LNCS, vol. 10961, pp. 763–772 (2018)
34. Balucani, N., et al.: A computational study of the reaction N(2D) + C6H6 leading to pyridine and phenylnitrene. In: LNCS, vol. 11621, pp. 316–324 (2019)
35. Balucani, N., et al.: An experimental and theoretical investigation of the N(2D) + C6H6 (benzene) reaction with implications for the photochemical models of Titan. Faraday Discuss. **245**, 327351 (2023). https://doi.org/10.1039/D3FD00057E
36. Rosi M., et al.: A computational study on the attack of nitrogen and oxygen atoms to toluene. In: LNCS, vol. 12953, pp. 620–631 (2021)
37. Rosi, M., et al.: A computational study on the insertion of N(2D) into a C—H or C—C bond: the reactions of N(^{2}D) with benzene and toluene and their implications on the chemistry of titan. In: LNCS, vol. 12251, pp. 744–755 (2020)
38. Rosi, M., et al.: A computational study of the reaction between N(^{2}D) and simple aromatic hydrocarbons. In: LNCS, vol. 14105, pp. 718–734 (2023)
39. Becke, A. D.: Density functional thermochemistry. III. The role of exact exchange. J. Chem. Phys. **98**, 5648–5652 (1993)
40. Stephens, P.J., Devlin, F.J., Chablowski, C.F., Frisch, M.J.: Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. J. Phys. Chem. **98**, 11623–11627 (1994)
41. Krishnan, R., Binkley, J.S., Seeger, R., Pople, J.A.: Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. J. Chem. Phys. **72**, 650–654 (1980)
42. Frisch, M.J., Pople, J.A., Binkley, J.S.: Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. J. Chem. Phys. **80**, 3265–3269 (1984)
43. Grimme, S., Ehrlich, S., Goerigk, L.: DFT-D3 - a dispersion correction for density functionals, Hartree-Fock and semi-empirical quantum chemical methods DFT-D3. J. Chem. Phys. **132**, 154104 (2010)
44. Dunning, Jr. T.H.: Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. J. Chem. Phys. **90**, 1007–1023 (1989)
45. Gonzalez, C., Schlegel, H.B.: An improved algorithm for reaction path following. J. Chem. Phys. **90**, 2154–2161 (1989)
46. Gonzalez, C., Schlegel, H.B.: Reaction path following in mass-weighted internal coordinates. J. Phys. Chem. **94**, 5523–5527 (1990)

47. Bartlett, R.J.: Many-body perturbation theory and coupled cluster theory for electron correlation in molecules. Annu. Rev. Phys. Chem. **32**, 359–401 (1981)
48. Raghavachari, K., Trucks, G.W., Pople, J.A., Head-Gordon, M.: Quadratic configuration interaction. A general technique for determining electron correlation energies. Chem. Phys. Lett. **157**, 479–483 (1989)
49. Olsen, J., Jorgensen, P., Koch, H., Balkova, A., Bartlett, R.J.: Full configuration–interaction and state of the art correlation calculations on water in a valence double-zeta basis with polarization functions. J. Chem. Phys. **104**, 8007–8015 (1996)
50. Moore, C.E.: Atomic energy levels. Natl. Bur. Stand. (U.S.) Circ. N. 467 (U.S., GPO, Washington, DC (1949)
51. Liang, P., et al.: OH ($^2\Pi$)+ C_2H_4 reaction: a combined crossed molecular beam and theoretical study. J. Phys. Chem. A **127**, 4609–4623 (2023)
52. Rosi, M., et al.: Electronic structure and kinetics calculations for the Si^+ SH reaction, a possible route of SiS formation in star-forming regions. In: LNCS, vol. 11621, pp. 306–315 (2019)
53. Mancini, L., et al.: A computational characterization of the reaction mechanisms for the reactions N (^{2}D)+ CH_3CN and HC_3N and implications for the nitrogen-rich organic chemistry of Titan. Comput. Theor. Chem. **1229**, 114341 (2023)
54. Mancini, L., et al.: Destruction of interstellar methyl cyanide (CH_3CN) via collisions with $He^{+\cdot}$ ions. Astron. Astrophys. **691**, A83 (2024)
55. Richardson, V., et al.: Fragmentation of interstellar methanol by collisions with He^+: an experimental and computational study. PCCP **24**(37), 22437–22452 (2022)
56. Valença Ferreira de Aragão, E., Mancini, L., Faginas-Lago, N., Rosi, M., Skouteris, D., Pirani, F.: Semiempirical potential in kinetics calculations on the HC_3N+ CN reaction. Molecules **27**(7), 2297 (2022)
57. Martin, J.M.: Ab initio total atomization energies of small molecules — towards the basis set limit. Chem. Phys. Lett. **259**, 669–678 (1996)
58. Svensson, M., Humbel, S., Froese, R.D.J., Matsubara, T., Sieber, S., Morokuma, K.: ONIOM: a multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels-Alder reactions and $Pt(P(t\text{-}Bu)_3)_2$ + H_2 oxidative addition. J. Phys. Chem. **100**, 19357 (1996)
59. Humbel, S., Sieber, S., Morokuma, K.: The IMOMO method: integration of different levels of molecular orbital approximations for geometry optimization of large systems: test for n-butane conformation and SN2 reaction: RCl+Cl^-. J. Chem. Phys. **105**, 1959 (1996)
60. Maseras, F., Morokuma, K.: IMOMM: a new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. J. Comp. Chem. **16**, 1170 (1995)
61. Frisch, M. J., et al.: Gaussian 09, Revision A.02. Gaussian, Inc., Wallingford CT (2009)
62. Werner, H.-J., Knowles, P.J., Knizia, G., Manby, F.R., Schütz, M.: Molpro: a general-purpose quantum chemistry program package. WIREs Comput. Mol. Sci. **2**, 242–253 (2012)
63. Werner, H.-J., et al.: The Molpro quantum chemistry package. J. Chem. Phys. **152**, 144107 (2020)
64. Flükiger, P., Lüthi, H.P., Portmann, S., Weber, J., MOLEKEL 4.3, Swiss Center for Sci-entific Computing, Manno (Switzerland), 2000–2002
65. Portmann, S., Lüthi, H. P., MOLEKEL: an interactive molecular graphics tool Chimia **54**, 766–769 (2000)
66. Mancini, L., Balucani, N., Rosi, M.: The reaction of N(^{2}D) with monocyclic and polycyclic aromatic hydrocarbons and their role in the atmospheric chemistry of Titan. Icarus (in press)

Can OSiS Be Formed in the Interstellar Medium? A Computational Investigation of the OH + SiS Reaction with a Kinetic Analysis

Nadia Balucani[1(✉)], Luca Mancini[1], Dimitrios Skouteris[2], Cecilia Ceccarelli[3], Linda Podio[4], Claudio Codella[4], Gabriella Di Genova[1], and Marzio Rosi[5]

[1] Dipartimento di Chimica, Biologia e Biotecnologie, Università degli Studi di Perugia, 06123 Perugia, Italy
{nadia.balucani,luca.mancini}@unipg.it
[2] Master-Tec SrL, Via Gerardo Dottori, 94, 06132 Perugia, Italy
[3] Université Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France
[4] INAF, Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Florence, Italy
[5] Dipartimento di Ingegneria Civile ed Ambientale, Università degli Studi di Perugia, 06125 Perugia, Italy
marzio.rosi@unipg.it

Abstract. A few years ago, silicon oxysulfide, OSiS, was suggested as a possible interstellar molecule containing both sulfur and silicon. Its detection using remote spectroscopy has so far been unsuccessful, and only upper limits have been reported in the literature. It remains to be seen whether efficient formation pathways exist that warrant further search. An initial suggestion for the OSiS formation indicated the O + SiS reaction, but this cannot form OSiS under the rarefied conditions of the interstellar medium, since stabilizing secondary or three-body collisions are highly unlikely. For this reason, following the example of reactions such as OH + CO, OH + CS, and OH + SiO we report a computational investigation of the OH + SiS reaction. According to our electronic structure calculations, which allowed us to construct the potential energy surface of the reactive system, the OH + SiS reaction can lead to the formation of OSiS + H, but a much more competitive pathway leads to SiO + HS. Our preliminary kinetic analysis showed that the yield of OSiS + H is ca. 17% at the temperatures of interest for the interstellar medium. Our results will be compared with those of a recent similar study, and the astrophysical implications will be noted.

Keywords: Electronic structure calculations · Kinetic calculations · Astrochemistry · Silicon chemistry

1 Introduction

Interstellar chemistry plays a fundamental role in the matter evolution within a galaxy, linking stellar nucleosynthesis to solar-system formation. Silicon and sulfur are both alpha process elements [1] and are among the most abundant elements of the third

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 461–473, 2027.
https://doi.org/10.1007/978-3-032-30521-3_29

period of the Periodic Table [2]. They are both believed to become part of refractory materials when the density is high and the temperature is low [3]. However, while there is little doubt that silicon is mostly retained as silicates in the core of interstellar dust grains [4], the main reservoir of interstellar sulfur is still debated [5]. In recent years, considerable effort has been made to characterize the abundance of gaseous S-bearing (see, for instance, [5–8]) and the possible inventory of S-molecules on ice (only $OCS_{(ice)}$ has been detected so far [9], and possibly $SO_{2(ice)}$ [10]). As of today, the puzzle of sulfur depletion in cold clouds and star-forming regions is far from being solved [5].

Silicon is heavily depleted onto dust grains under quiescent conditions. Therefore, the presence of its molecules in the gas phase is considered an important tracer of shock-driven chemistry, as energetic processes are required to release refractory elements from dust grains. Molecules containing both sulfur and silicon, such as silicon sulfide (SiS), may offer insights into the long-standing problem of sulfur depletion in dense clouds since both S and Si are released in the gas phase by shocks and outflows [11]. Numerous formation routes of SiS have been considered, including neutral-neutral reactions [12–17] and ion-molecule reactions [18,19]. In addition to SiS, silicon oxysulfide, OSiS, has been proposed as another possible interstellar molecule where both sulfur and silicon are present [20]. After its isolation and characterization in a cold-matrix experiment [21], the rotational spectroscopy of OSiS was characterized in the gas phase [22], and its astronomical search has been pursued to date without success [23].

Even before confirming its presence (only upper limits are available in the literature [23]), the possible formation routes of OSiS have been analyzed. In shocked regions, SiS is normally present, and therefore, the formation pathways of OSiS could start from SiS. An initial study considered the O + SiS reaction [20], but it cannot work in the rarefied conditions of the interstellar medium because stabilizing secondary or three-body collisions are highly improbable. Considering the analogous reactions OH + CO → CO_2 + H [24], OH + SiO → SiO_2 + H [25], and OH + CS → OCS + H [26], the OH + SiS reaction, which is the subject of this study, emerges as a possible formation route of OSiS. This reaction was considered by Costa et al. [27] very recently, together with other silicon reactions. They characterized the OH + SiS potential energy surface (PES) at the CCSD(T)-F12/cc-pVQZF12//M06-2X/aug-cc-pV(T+d)Z + ZPE(M06-2X/aug-ccpV(T+d)Z) level of theory.

Here, we present our calculations obtained at the CCSD(T)/CBS//ωB97XD/aug-cc-pV(T+d)Z) level of theory. In addition to deriving a more complete PES, in which additional pathways are present, we have also performed Rice-Ramsperger-Kassel-Marcus (RRKM) calculations to estimate the product branching fractions of the two open channels

$$\mathrm{OH} + \mathrm{SiS} \longrightarrow \mathrm{HS} + \mathrm{SiO} \tag{1}$$

$$\mathrm{OH} + \mathrm{SiS} \longrightarrow \mathrm{OSiS} + \mathrm{H}. \tag{2}$$

The computed branching fraction of reaction (2) is smaller than the value estimated by Costa et al. [27] without performing a kinetic analysis of the PES.

2 Theoretical Methods

The OH + SiS reaction was characterized by adopting the same computational strategy already used for other bimolecular reactions (see, for instance, [12,28–35]) and using the GAUSSIAN09 [36] and Molpro [37] software packages. All stationary points were first optimized at the DFT ωB97XD level of theory [38], in conjunction with the correlation consistent valence polarised set aug-cc-pV(T+d)Z [39–41]. Harmonic vibrational frequencies were computed, at the same level of theory, to determine the nature of each stationary point, that is, a minimum (reactants, reaction intermediates and products) if all the frequencies are real, and a first-order saddle point (transition state) if there is one, and only one, imaginary frequency. When a saddle point was found, Intrinsic Reaction Coordinate (IRC) calculations [42,43] were performed to assign each transition state to the respective reactant and product. After each geometry optimization at the DFT level, more accurate calculations were performed using the single- and double-electronic excitation coupled-cluster method with perturbative description of triple excitations, CCSD(T) [44–46], corrected with a Density Fitted (DF)MP2 extrapolation to the complete basis set (CBS) and with corrections for core electron excitations. Namely, the energy values were computed according to:

$$E(\text{CCSD(T)/aug-cc-pVTZ}) + E(1) + E(2) \tag{3}$$

where

$$E(1) = E(\text{CCSD(T,core)/cc-pVTZ}) - E(\text{CCSD(T)/cc-pVTZ})$$
$$E(2) = E(\text{DF-MP2/CBS}) - E(\text{DF-MP2/aug-cc-pVTZ})$$

with E(DF-MP2/CBS) defined as:

$$\begin{aligned} E(\text{DF-MP2/CBS}) = {} & E[(\text{DF-MP2})/\text{aug-cc-pVQZ})] \\ & + 0.5772 * [E(\text{DF-MP2/aug-cc-pVQZ}) \\ & - E(\text{DF-MP2/aug-cc-pVTZ})] \end{aligned} \tag{4}$$

The E(DF-MP2/CBS) extrapolation was performed using Martin's two-parameter scheme [47], leading to an expected accuracy of about ± 5 kJ mol^{-1}. These energies, labeled CBS despite also including core-valence correlation corrections, were used for the kinetic investigation described below. The zero-point-energy (ZPE), evaluated at the DFT level of theory, was used to correct all energy values. Given the uncertainty in the calculations, all reported energy values have been rounded to the nearest integer. In a few cases, an additional decimal place is provided to indicate the relative ordering of structures with very similar energies.

The branching fraction of the products has been calculated by employing a combination of RRKM and master-equation programs developed by us and previously described in detail [48–50].

3 Results

The complete PES is reported in Fig. 1, while the structures of the reactants, products, intermediates, and transition states are shown in Figs. 2, 3, and 4. The energy values

of the stationary points obtained at the two employed levels of theory are shown in Fig. 1 (black: DFT ωB97XD level of theory; in blue and in brackets the data obtained at the CBS level) and reported in Table 1. To evaluate the possible multireference character of the investigated sulfur-containing systems, T1 diagnostics were computed for the relevant stationary points [51]. The obtained values (ca. 0.02, consistently below the accepted threshold of 0.03) indicate that multireference effects are limited, supporting the suitability of the present approach for reliably describing the present potential energy surfaces.

The reaction starts with the addition of OH to SiS, either on the S or Si side.

When adding to the Si side, the addition intermediate MIN1, quite stable with respect to the reactants (−284 kJ/mol at the CBS level of theory), is formed. Once formed, MIN1 can dissociate directly via the O-H bond fission and form OSiS + H. However, to do that, TS2 (a very high-energy transition state, +254 kJ/mol) must be overcome, making this direct channel quite improbable. Instead, MIN1 can easily isomerize into MIN2c by overcoming a much lower transition state (TS1, located at −121 kJ/mol) or into MIN3 by overcoming TS5 (−102 kJ/mol). Once formed, MIN2c can, in turn, easily isomerize to its *trans* isomer MIN2t (via TS7 located at −204 kJ/mol) or dissociate into SiO + SH (a channel without an exit barrier) or into OSiS + H; in this case, an energy barrier associated with TS3 (−45 kJ/mol) must be overcome. Also MIN2t can dissociate into SiO + SH via a barrierless exit channel or into OSiS + H by overcoming an exit barrier associated with TS8 (−43 kJ/mol). Given the very similar geometries and energies, TS3 and TS8 may be regarded as conformationally related transition structures associated with distinct cis/trans intermediates. MIN3 can only isomerize back to MIN1 or dissociate into OSiS + H by passing through the critical configuration of TS6. This transition state is located at −40 kJ/mol at the DFT ωB97XD level of theory (and in this case it is an exit barrier toward the products OSiS + H, whose energy content is −43 kJ/mol) or at −57 kJ/mol at the CBS level of theory. In this second case, TS6 exhibits the characteristics of a submerged transition state along the route toward the OSiS + H product (characterized by an energy content of −55 kJ/mol at the same level of theory).

When adding to the S side, the MIN4c intermediate is formed, associated with a swallow minimum (−43 kJ/mol at the CBS level). MIN4c can easily isomerize to its *trans* isomer MIN4t (−41 kJ/mol at the CBS level) via TS9. MIN4t can isomerize to MIN1 by overcoming a barrier associated with TS4 located at −5 kJ/mol relative to the reactants. Therefore, TS4 is a submerged transition state at this level of theory. Given the uncertainty in the current calculations, TS could also be at the same energy level as, or slightly above, that of the reactants.

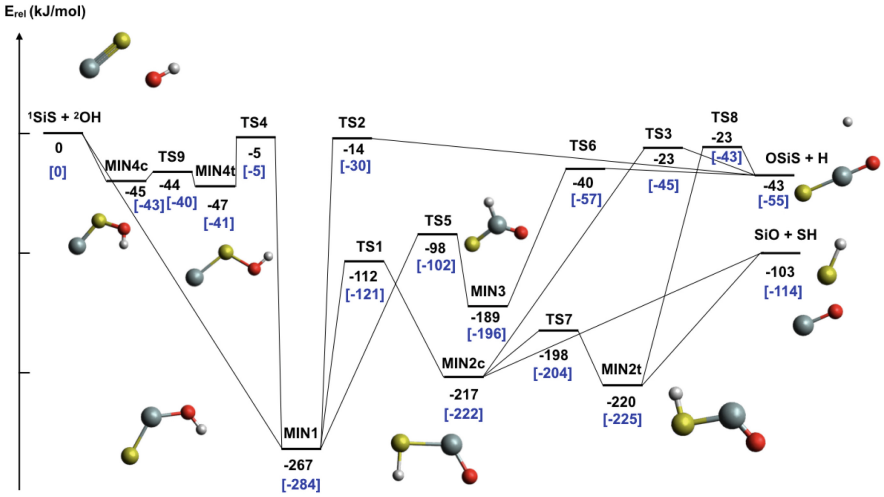


Fig. 1. Schematic representation of the potential energy surface obtained for the OH + SiS reaction at the CCSD(T)/CBS//ωB97XD/aug-cc-pV(T+d)Z) level of theory. Sulfur atoms are shown in yellow, silicon in gray, oxygen in red, and hydrogen in white. (Color figure online)

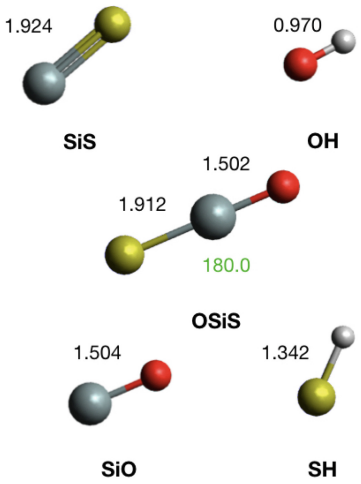


Fig. 2. ωB97XD optimized geometries (Å and deg) of the reactants and products. Sulfur atoms are shown in yellow, silicon in gray, oxygen in red, and hydrogen in white. (Color figure online)

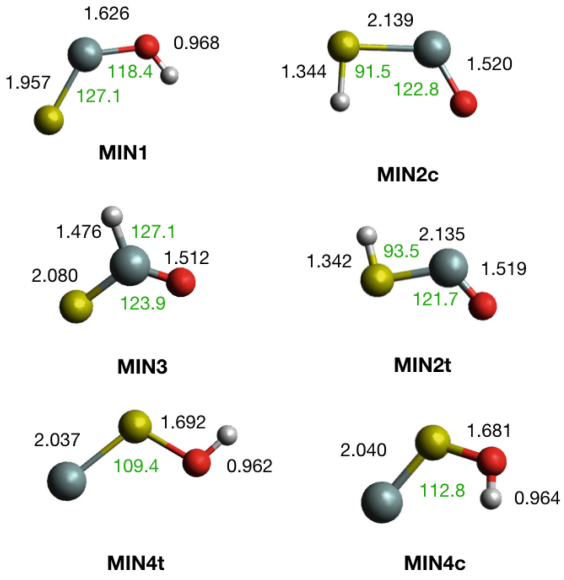


Fig. 3. ωB97XD optimized geometries (Å and deg) of the intermediates formed during the reaction OH + SiS (see Fig. 1). Sulfur atoms are shown in yellow, silicon in gray, oxygen in red, and hydrogen in white.

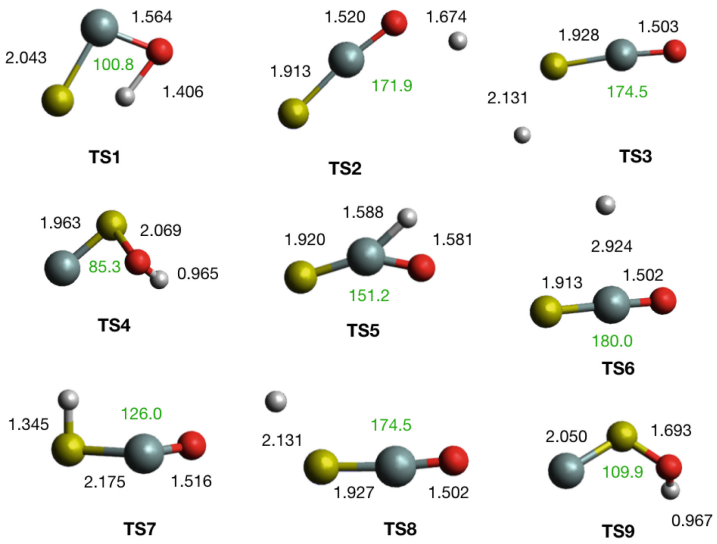


Fig. 4. ωB97XD optimized geometries (Å and deg) of the transition states characterizing the OH + SiS PES (see Fig. 1). Sulfur atoms are shown in yellow, silicon in gray, oxygen in red, and hydrogen in white.

Our PES scheme has many similarities to the one derived by Costa et al. [27], but also a few differences. For example, the MIN4 intermediates cannot evolve along their PES, whereas in our case they isomerize to MIN1 by overcoming a submerged barrier (TS4). Table 1 reports the energy values of the stationary points of the PES calculated in this work at the ωB97XD/aug-cc-pV(T+d)Z and CCSD(T)/CBS//ωB97XD/aug-cc-pV(T+d)Z) level of theory to be compared with those reported by Costa et al., obtained at the CCSD(T)-F12/cc-pVQZF12//M06-2X/aug-cc-pV(T+d)Z level of theory. In many cases, the small differences found between our CBS calculations and theirs are within the uncertainty of the methods used.

The most significant differences in the two PESs is actually related to the reverse reaction H + OSiS. Indeed, in our PES there is a significant exit barrier along the possible pathways leading to OSiS + H, see the sequences MIN1 → TS2 → OSiS+H, MIN2c → TS3 → OSiS+H, MIN2t → TS8 → OSiS+H. The structures of the involved TS are shown in Fig. 4. The only pathway in which the TS associated with the formation of the products is submerged is MIN3 → TS6 → OSiS+H. According to the calculations by Costa et al., instead, the pathways i3 → OSiS+H (equivalent to our MIN3 → TS6 → OSiS+H) and i4 → OSiS+H (equivalent to our MIN2t → TS8 → OSiS+H) are barrierless. In the case of MIN3 → TS6 → OSiS+H, the energy of TS6 is −40 kJ/mol at the ωB97XD/aug-cc-pV(T+d)Z level to be compared with ΔH_0^o(OSiS + H) of −43 kJ/mol but, if we consider the CBS higher level of calculations, TS6 is a submerged transition state having an energy of −57 kJ/mol to be compared with ΔH_0^o(OSiS + H) of −55 kJ/mol. In other words, while our work excludes that the other H attacks on OSiS are barrierless, in this specific case, the formation of MIN3 from the H + OSiS reaction could also occur under cold-cloud conditions in the interstellar medium. Since hydrogen atoms are ubiquitous in the interstellar medium, the fact that the reaction can occur without a barrier in the entrance channel is extremely relevant to determine the resistance of OSiS, since the reaction with H would very easily destroy it. A similar condition had already been discussed for the H + SiS_2 reaction [12,49]: the lack of a barrier in the entrance channel makes the reaction extremely fast, systematically converting interstellar SiS_2 (that can be formed by the SiH + S_2 reaction) into SiS+HS. Remarkably, also the H + SiO_2 reaction is not characterized by an entrance barrier. However, in that case, there are no open exothermic channels and SiO_2, once formed, cannot be destroyed by the H atoms (see the Discussion in [49]). To be noted that also in the case of H + SiO_2 and H + SiS_2 it is the formation of trigonal structures with the Si engaged in three bonds that can be formed without an activation energy, while the H addition to one of the O atoms of SiO_2 or one of the S atoms of SiS_2 is always associated with an entrance barrier. This result is somewhat surprising because of the available lone pairs on O and S atoms. Furthermore, the addition of H to the central Si atom of OSiS, SiS_2, and SiO_2 requires a hybridization change of the central Si atom from sp to sp^2. Clearly, the new Si-H bond is strong enough to compensate for the significant change required in the molecular backbone.

3.1 The Product Branching Fraction

Since this calculation serves only to establish the branching fraction between the two channels (1) and (2), we did not conduct a complete kinetic analysis of the PES

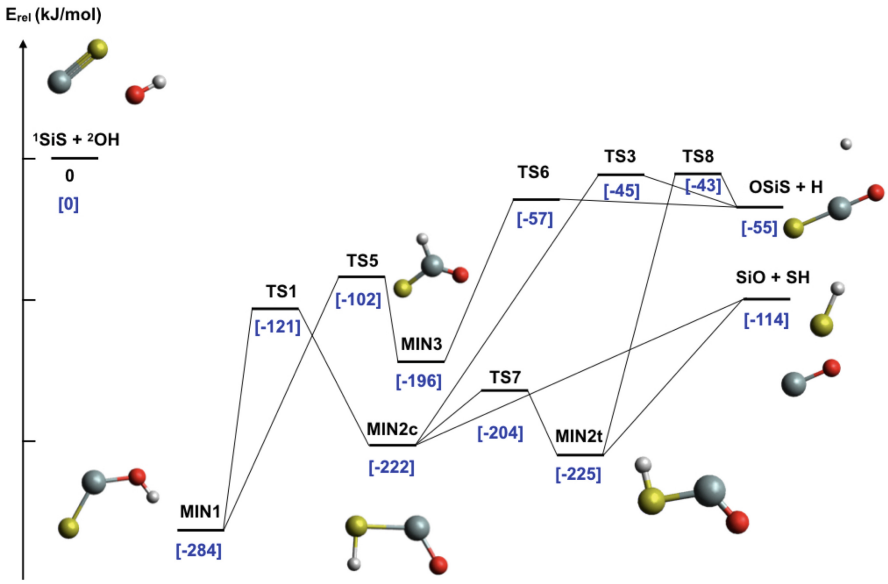


Fig. 5. Portion of the potential energy surface illustrated in Fig. 1 that has been used to evaluate the product branching fractions via RRKM calculations. Sulfur atoms are shown in yellow, silicon in gray, oxygen in red, and hydrogen in white.

(deferred to more comprehensive future work). We started from MIN1 and considered only the pathways originating after its isomerization to MIN2c and MIN3, i.e., we have analyzed only the reduced portion of the PES shown in Fig. 5. After solving the master equation for the paths illustrated in Fig. 5 and averaging the microcanonical constants over the Boltzmann distribution, we calculated the BFs at the most interesting temperatures in the interstellar medium, namely 10, 100, and 200 K (see Table 2. We also calculated the value at room temperature. In line with the characteristics of the PES, the SiO + HS formation channel is by far the dominant one and there is a little trend with the temperature with the yield of OSiS + H iscreasing from 16.6% to 18.4%. Therefore, the yield of OSiS is lower than the 25% yield suggested by Costa et al., who assumed this value at all temperatures considered.

Table 1. Energies (kJ/mol, 0 K) of the stationary points of the OH + SiS PES, computed at the indicated level of theory (present work: ωB97XD/aug-cc-pV(T+d)Z and CCSD(T)/CBSωB97XD/aug-cc-pV(T+d)Z) level of theory; Costa et al. [27]: CCSD(T)-F12/cc-pVQZF12//M06-2X/aug-cc-pV(T+d)Z level of theory).

	ωB97XD/aug-cc-pV(T+d)Z this work	CBS this work	CCSD(T)- F12/cc-pVQZF12// M06-2X/aug-cc-pV(T+d)Z [27]
MIN1	267	−284	−281 (i2c in [27])
MIN2(cis)	−217	−222	−218
MIN2(trans)	−220	−225	−222 (i4t in [27])
MIN3	−189	−196	−190 (i3 in [27])
MIN4(cis)	−45	−43	−39 (i1c in [27])
MIN4(trans)	−47	−41	−37
TS1	−112	−121	−117
TS2	−14	−30	−30
TS3	−23	−45	/
TS4	−5	−5	/
TS5	−98	−102	−98
TS6	−40	−57	/
TS7	−198	−204	/
TS8	−23	−43	/
TS9	−44	−40	/
ΔH_0^o(SiO + SH)	−103	−114	−110
ΔH_0^o(OSiS + H)	−43	−55	−54

4 Discussion and Astrophysical Implications

In this work, we addressed the theoretical study of the OH + SiS reaction as a possible mechanism for the formation of OSiS in shocked regions of the interstellar medium. Based on other similar reactions, particularly OH + CS, OH + SiO, and OH + CO, it was reasonable to expect the formation of OSiS. However, potential energy surface analysis immediately shows that the reaction channel leading to OSiS + H is strongly disfavored compared to the more exothermic SiO + HS channel. Kinetic analysis using the RRKM method confirmed a lower yield of channel (2) (<18% under conditions typical of star-forming regions). Therefore, the value suggested by Costa et al. [27] overestimates the yield of channel (2).

Furthermore, the previous conclusions by Costa et al. [27] on the destruction of OSiS by atomic hydrogen require further investigation. It is unclear whether or not an activation energy exists for the reaction H + OSiS → SiO + HS. However, even if TS6 is confirmed to be submerged relative to H + OSiS, its presence still represents a bottleneck for the reactive flux. Since there is certainly a van der Waals complex between the reactants and TS6, the combination of the two will increase back-dissociation. Therefore, a gas kinetic rate constant such as the one suggested by Costa et al. (4×10^{-10}

Table 2. Product branching fractions at different temperatures of relevance to the chemistry of the interstellar medium. Calculations were performed using the RRKM method, taking into account the paths originating from the MIN2 intermediate, as these are the minimum energy paths for the formation of both the SiO + HS and OSiS + H products (see the scheme reported in Fig. 5).

channel	T = 10 K	T = 100 K	T = 200 K	T = 300 K
SiO + SH	83.4%	83.0%	82.1%	81.6%
OSiS + H	16.6%	17.0%	17.9%	18.4%

$cm^3\ mol^{-1}$) seems to be implausible. Clearly, the most important role of the OH + SiS reaction is to convert SiS to SiO as already observed in the reaction O + SiS [14].

5 Conclusions

This study represents another piece in the puzzle of SiS chemistry. There are many reasons to try to understand the genesis of this molecule in the interstellar medium and also in the circumstellar envelopes of asymptotic giant branch and red supergiants stars. Its relationship with the similar molecule SiO is not only linked to the larger abundance of oxygen compared to sulfur. Since both silicon and sulfur are strongly depleted in the dense regions of the interstellar medium, they are released together into the gas phase due to shock waves. SiO and SiS probably form at different times, and their fate, once formed, is also different. An excess of oxygen species systematically converts SiS to SiO. This was already known for the O + SiS reaction and is now also known for the OH + SiS reaction. The formation of OSiS by reactions involving SiS does not seem to contribute to preserving the Si-S bond because they are too uncompetitive with other reaction channels that convert SiS to SiO. There are still many aspects to be clarified and which require dedicated studies such as the one shown here.

Acknowledgments. The authors thank the Italian MUR PRIN2020 (2020AFB3FX - Astrochemistry beyond the second period elements). MR and LP acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title 2022JC2Y93 ChemicalOrigins: linking the fossil composition of the Solar System with the chemistry of protoplanetary disks – CUP J53D23001600006 - Grant Assignment Decree No. 962 adopted on 30.06.2023 by the Italian Ministry of Ministry of University and Research (MUR).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Goswami, S., Vilchez, J.M., Pérez-Díaz, B., Silva, L., Bressan, A., Pérez-Montero, E.: Contribution of very massive stars to the sulfur abundance in star-forming galaxies: role of pair-instability supernovae. A&A **685**, A81 (2024). https://doi.org/10.1051/0004-6361/202348231
2. Asplund, M., Grevesse, N., Sauval, A.J., Scott, P.: The chemical composition of the sun **47**(1), 481–522 (2009). https://doi.org/10.1146/annurev.astro.46.060407.145222
3. Jenkins, E.B.: A unified representation of gas-phase element depletions in the interstellar medium. Astrophys. J. **700**(2), 1299 (2009). https://doi.org/10.1088/0004-637X/700/2/1299
4. Draine, B.T.: Interstellar dust grains, vol. 41, pp. 241–289 (2003). https://doi.org/10.1146/annurev.astro.41.011802.094840
5. Laas, J.C., Caselli, P.: Modeling sulfur depletion in interstellar clouds. A&A **624**, A108 (2019). https://doi.org/10.1051/0004-6361/201834446
6. Fuente, A., et al.: Gas phase elemental abundances in molecular clouds (GEMS) - VII. sulfur elemental abundance. A&A **670**, A114 (2023). https://doi.org/10.1051/0004-6361/202244843
7. Rivière-Marichalar, P., et al.: Observations of sulfuretted species in HL Tau. A&A **703**, A282 (2025)
8. Taquet, V., et al.: Seeds of life in space (SOLIS) - VI. Chemical evolution of sulfuretted species along the outflows driven by the low-mass protostellar binary NGC 1333-IRAS4A. A&A **637**, A63 (2020). https://doi.org/10.1051/0004-6361/201937072
9. Palumbo, M., Geballe, T., Tielens, A.G.: Solid carbonyl sulfide (OCS) in dense molecular clouds. Astrophys. J. **479**(2), 839–844 (1997)
10. Gross, R.E., et al.: CORINOS. IV. Quantifying Baseline-fitting Uncertainties in SO_2 ice measurements with JWST/MIRI **998**(1), 68 (2026). https://doi.org/10.3847/1538-4357/ae2c81
11. Podio, L., et al.: Silicon-bearing molecules in the shock L1157–B1: first detection of SiS around a Sun-like protostar. Mon. Not. Roy. Astron. Soc. Lett. **470**(1), L16–L20 (2017)
12. Rosi, M., et al.: Possible scenarios for SiS formation in the interstellar medium: electronic structure calculations of the potential energy surfaces for the reactions of the SiH radical with atomic sulphur and S_2. Chem. Phys. Lett. **695**, 87–93 (2018)
13. Rosi, M., et al.: Electronic structure and kinetics calculations for the Si+SH reaction, a possible route of SiS formation in star-forming regions. In: Misra, S., et al. (eds.) ICCSA 2019. LNCS, vol. 11621, pp. 306–315. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-24302-9_22
14. Zanchet, A., Roncero, O., Agundez, M., Cernicharo, J.: Formation and destruction of SiS in space. ApJ **862**(1), 38 (2018). https://doi.org/10.3847/1538-4357/aaccff
15. Mota, V.C., Caridade, P.J.S.B., Varandas, A.J.C., Galvao, B.R.L.: Quasiclassical trajectory study of the SI plus SH reaction on an accurate double many-body expansion potential energy surface. J. Phys. Chem. A **126**(22), 3555–3568 (2022). https://doi.org/10.1021/acs.jpca.2c01633
16. Doddipatla, S., He, C., Goettl, S.J., Kaiser, R.I., Galvão, B.R., Millar, T.J.: Nonadiabatic reaction dynamics to silicon monosulfide (SiS): a key molecular building block to sulfur-rich interstellar grains. Sci. Adv. **7**(26), eabg7003 (2021)
17. Fortenberry, R.C., McGuire, B.A.: A possible additional formation pathway for the interstellar diatomic SiS. Astroph. J. **971**(1), 101 (2024). https://doi.org/10.3847/1538-4357/ad4d94
18. Michielan, M., et al.: Experimental and computational investigation of an ionic pathway to the formation of silicon sulphide in the interstellar medium. A&A **698**, A205 (2025)
19. Mancini, L., Trinari, M., de Aragão, E.V.F., Rosi, M., Balucani, N.: The S+(4 S)+ SiH 2 (1 A 1) reaction: toward the synthesis of interstellar SiS. In: International Conference on

Computational Science and Its Applications, pp. 233–245. Springer (2022). https://doi.org/10.1007/978-3-031-10562-3_17
20. Paiva, M.A.M., Lefloch, B., Galvão, B.R.L.: Accurate theoretical predictions on the mechanisms of OSiS formation and its dissociation channels. Mon. Not. Roy. Astron. Soc. **481**(2), 1858–1861 (2018). https://doi.org/10.1093/mnras/sty2382
21. Schnöckel, H.: Matrix isolation of oSiS: IR spectroscopic evidence for the sis double bond. Angew. Chem. Int. Ed. Engl. **19**(4), 323–324 (1980). https://doi.org/10.1002/anie.198003231
22. Thorwirth, S., Mück, L.A., Gauss, J., Tamassia, F., Lattanzi, V., McCarthy, M.C.: Silicon Oxysulfide, OSiS: rotational spectrum, quantum-chemical calculations, and equilibrium structure. J. Phys. Chem. Lett. **2**(11), 1228–1231 (2011). https://doi.org/10.1021/jz200368x
23. Esplugues, G.B., et al.: A line confusion-limited millimeter survey of Orion KL - III. Sulfur oxide species. A&A **556**, A143 (2013). https://doi.org/10.1051/0004-6361/201321285
24. Caracciolo, A., et al.: Combined experimental–theoretical study of the OH + CO → H + CO_2 reaction dynamics. J. Phys. Chem. Lett. **9**(6), 1229–1236 (2018). https://doi.org/10.1021/acs.jpclett.7b03439
25. Martín, J.C.G., Blitz, M.A., Plane, J.M.: Kinetic studies of atmospherically relevant silicon chemistry. Part II: Silicon monoxide reactions. Phys. Chem. Chem. Phys. **11**(46), 10945–10954 (2009)
26. Adriaens, D., Goumans, T., Catlow, C., Brown, W.: Computational study of carbonyl sulphide formation on model interstellar dust grains. J. Phys. Chem. C **114**(4), 1892–1900 (2010)
27. Costa, S.F., Alves, M.O., Pilling, S., Li, J., Galvão, B.R.: New neutral-neutral pathways in the chemistry of silicon-and sulfur-bearing molecules in the cold interstellar medium. Astrophys. J. **995**(2), 209 (2025)
28. Skouteris, D., Balucani, N., Faginas-Lago, N., Falcinelli, S., Rosi, M.: Dimerization of methanimine and its charged species in the atmosphere of titan and interstellar/cometary ice analogs. Astron. Astrophys. **584**, A76 (2015)
29. Giani, L., et al.: Revised gas-phase formation network of methyl cyanide: the origin of methyl cyanide and methanol abundance correlation in Hot Corinos. Mon. Not. R. Astron. Soc. **526**(3), 4535–4556 (2023)
30. Rosi, M., et al.: Electronic structure and kinetics calculations for the SI+ SH reaction, a possible route of sis formation in star-forming regions. In: International Conference on Computational Science and Its Applications, pp. 306–315. Springer (2019)
31. Liang, P., et al.: Oh (2π)+ C2H4 reaction: a combined crossed molecular beam and theoretical study. J. Phys. Chem. A **127**(21), 4609–4623 (2023)
32. Mancini, L., et al.: A computational characterization of the reaction mechanisms for the reactions n (2D)+ CH3CN and HC3N and implications for the nitrogen-rich organic chemistry of titan. Comput. Theor. Chem. **1229**, 114341 (2023)
33. Mancini, L., et al.: Destruction of interstellar methyl cyanide (CH3CN) via collisions with HE+ ions. Astron. Astrophys. **691**, A83 (2024)
34. Richardson, V., et al.: Fragmentation of interstellar methanol by collisions with HE+: an experimental and computational study. Phys. Chem. Chem. Phys. **24**(37), 22437–22452 (2022)
35. Valença Ferreira de Aragão, E., Mancini, L., Faginas-Lago, N., Rosi, M., Skouteris, D., Pirani, F.: Semiempirical potential in kinetics calculations on the HC3N+ CN reaction. Mol. **27**(7), 2297 (2022)
36. Frisch, M., et al.: Fox: Gaussian 09, Revision A. 02, 2009, Gaussian. Inc., Wallingford CT (2009)
37. Werner, H.J., et al.: The Molpro quantum chemistry package. J. Chem. Phys. **152**(14), 144107 (2020)

38. Chai, J.D., Head-Gordon, M.: Systematic optimization of long-range corrected hybrid density functionals. J. Chem. Phys. **128**(8) (2008)
39. Dunning, Jr. T.H.: Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. J. Chem. Phys. **90**(2), 1007–1023 (1989)
40. Kendall, R.A., Dunning, Jr. T.H., Harrison, R.J.: Electron affinities of the first-row atoms revisited. systematic basis sets and wave functions. J. Chem. Phys. **96**(9), 6796–6806 (1992)
41. Woon, D.E., Dunning, Jr. T.H.: Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. J. Chem. Phys. **98**(2), 1358–1371 (1993)
42. Gonzalez, C., Schlegel, H.B.: An improved algorithm for reaction path following. J. Chem. Phys. **90**(4), 2154–2161 (1989). https://doi.org/10.1063/1.456010
43. Gonzalez, C., Schlegel, H.B.: Reaction path following in mass-weighted internal coordinates. J. Phys. Chem. **94**(14), 5523–5527 (1990). https://doi.org/10.1021/j100377a021
44. Bartlett, R.J.: Many-body perturbation theory and coupled cluster theory for electron correlation in molecules. Annu. Rev. Phys. Chem. **32**(1), 359–401 (1981)
45. Olsen, J., Jo/rgensen, P., Koch, H., Balkova, A., Bartlett, R.J.: Full configuration–interaction and state of the art correlation calculations on water in a valence double-zeta basis with polarization functions. J. Chem. Phys. **104**(20), 8007–8015 (1996)
46. Raghavachari, K., Trucks, G.W., Pople, J.A., Head-Gordon, M.: A fifth-order perturbation comparison of electron correlation theories. Chem. Phys. Lett. **157**(6), 479–483 (1989)
47. Martin, J.M.: Ab initio total atomization energies of small molecules–towards the basis set limit. Chem. Phys. Lett. **259**(5–6), 669–678 (1996)
48. Balucani, N., et al.: Combined crossed molecular beam and theoretical studies of the N(^{2}D)+ CH_4 reaction and implications for atmospheric models of titan. J. Phys. Chem. A **113**(42), 11138–11152 (2009). https://doi.org/10.1021/jp904302g
49. Skouteris, D., et al.: A theoretical investigation of the reaction H+SiS2 and implications for the chemistry of silicon in the interstellar medium. In: Gervasi, O., et al. (eds.) ICCSA 2018. LNCS, vol. 10961, pp. 719–729. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95165-2_50
50. Di Genova, G., et al.: Gas-phase formation routes of dimethyl sulfide in the interstellar medium. Astron. Astrophys. **704**, A305 (2025)
51. Lee, T.J., Taylor, P.R.: A diagnostic for determining the quality of single-reference electron correlation methods. Int. J. Quantum Chem. **36**(S23), 199–207 (1989)

Workshop on Computational Science and HPC (CSHPC 2026)

Numerical Evaluation of 3-Loop "Baseball" and "Running Shirt" Feynman Diagrams Using GPU-Accelerated QMC Integration and Extrapolation, Based on Sector Decomposition

Elise de Doncker[1(✉)], Hiroshi Daisaka[3], Tadashi Ishikawa[2], Kiyoshi Kato[4], Naohito Nakasato[5], and Fukuko Yuasa[2]

[1] Western Michigan University, Kalamazoo, MI 49008, USA
elise.dedoncker@wmich.edu
[2] High Energy Accelerator Research Organization (KEK), 1-1 OHO, Tsukuba, Ibaraki 305-0801, Japan
{tadashi.ishikawa,fukuko.yuasa}@kek.jp
[3] Graduate School of Business Administration, Hitotsubashi University, Kunitachi, Tokyo 186-8601, Japan
daisaka@phys.science.hit-u.ac.jp
[4] Department of Physics, Kogakuin University, Shinjuku, Tokyo 163-8677, Japan
katok@kute.tokyo
[5] Department of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu, Fukushima 965-8580, Japan
nakasato@u-aizu.ac.jp

Abstract. We consider integrals associated with two-point Feynman diagrams that have three loops and six internal lines. In particular, this paper studies two diagrams that, in view of their shape, are named "Baseball" and "Running Shirt" diagrams. Since the integrals are generally divergent, we transformed the integral into a suitable form using the sector decomposition method, and evaluated the resulting integrals with singular integrands using appropriate numerical methods.The combination of the lattice Quasi-Monte Carlo (QMC) rule and the extrapolation is used for the computation. The integration is accelerated using GPUs.

Keywords: 3-loop Feynman diagrams · QMC integration · Extrapolation · GPU

1 Introduction

Improvements in the technology of high energy physics experiments require theoretical predictions of commensurate accuracy. Higher order corrections are needed for accurate theoretical predictions of the cross-section for particle interactions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 477–491, 2027.
https://doi.org/10.1007/978-3-032-30521-3_30

The Feynman diagrammatic approach is commonly used to address higher order corrections, and Feynman loop integrals arise in the calculations. Symbolic or symbolic/numerical calculations are done for some challenging problems using established software packages such as pySecDec [4], FIESTA [25] and MBnumerics [26].

A general and analytic treatment is not possible for general mass and momentum values, so we resort to numerical approaches. The number of dimensions and the degree of difficulty of the integration increase as the number of lines and loops increases, emphasizing the use of multivariate integration methods.

A Feynman integral with 3 loops and 6 lines can be represented by

$$\mathbf{I} = (-1)^6 \frac{\Gamma(6-3\nu/2)}{(4\pi)^{3\nu/2}} \int_{\mathcal{K}_6} \prod_{r=1}^{6} dx_r \frac{\delta(1-\sum x_r)}{U^{\nu/2} V^{6-3\nu/2}} = \frac{\Gamma(1+3\varepsilon)}{(4\pi)^{6-3\varepsilon}} \frac{1}{3\varepsilon} I. \qquad (1)$$

where U and V are determined by the topology of the corresponding diagram and physical parameters; $\nu = 4 - 2\varepsilon$ is the space-time dimension; $\mathcal{K}_6$ is the 6-dimensional unit hypercube. V for a two-point function is written in terms of a W function as

$$V = M^2 - s\frac{W}{U}, \quad M^2 = \sum_{r=1}^{6} m_r^2 x_r - i\varrho \qquad (2)$$

where the m_r are the particle masses of internal lines and s is the squared momentum P that flows into the diagram. The U and W functions are polynomials of Feynman parameters, x_r; U is third order and W is of order 4. As shown in Eq. (3), U is semi-positive definite and the zero occurring at the endpoint of the variables can produce the ultraviolet (UV) singularity, which is extracted efficiently through the sector decomposition method. V is positive for $s < 0$, and when we increase the value of s, it becomes zero in $\mathcal{K}_6$ above a certain value of s which we call the threshold. ϱ is a parameter to alleviate the singular behavior above the threshold.

There are eleven diagrams contributing to the 3-loop two-point function with six internal lines when the model for the Feynman diagram generation is restricted to include up to five-point vertices and to diagrams that are neither products of lower-order functions nor energy-independent [17]. Among these eleven diagrams, we already presented the result for the one diagram free from UV singularities in [10], as well as the results for the two diagrams involving the "complete" four-variable U function in [6].

In this paper, we handle the " Baseball (BB)" and " Running Shirt (RS)" diagrams as case studies (cf., Fig. 1). It is worth noting that diagrams (2n) and (2m) in Laporta [15] give rise to the same integrals as BB and RS, respectively.

Via a transformation $x_j = y_j$, $(1 \le j \le 4)$, $x_5 = y_5 z$, $x_6 = y_5 \bar{z}$, the function U for these diagrams is rewritten as

$$(\mathrm{BB}, \mathrm{RS}) \qquad U = y_1y_2y_3 + y_1y_2y_4 + y_1y_3y_4 + y_2y_3y_4 + y_{12}y_{34}y_5, \qquad (3)$$

and the W functions are rewritten as

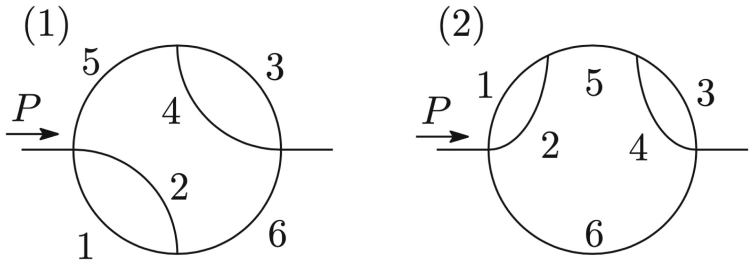


Fig. 1. Baseball (1) and Running Shirt (2) Feynman diagrams. The numbers assigned to the propagator denote the index r in Eq. (2)

$$\text{(BB)} \qquad W = y_1y_2y_3y_4 + y_1y_2y_{34}y_5z + y_{12}y_3y_4y_5\bar{z} + y_{12}y_{34}y_5^2z\bar{z}\,, \tag{4}$$

$$\text{(RS)} \qquad W = (y_1y_2y_{34} + y_{12}y_3y_4)y_5\bar{z} + y_{12}y_{34}y_5^2z\bar{z}\,. \tag{5}$$

where we use the notation $y_{ij} = y_i + y_j$ and $\bar{z} = 1 - z$. In this paper, we consider the case where all masses are equal ($m_r = m$).

While partial results for BB were presented in our previous work [7,9] (for only one or two sectors), in this paper we complete the evaluation of all contributions (15 sectors) to both BB and RS and present the results. In other work, we handled diagrams with up to five loops (e.g., [6–9,11]).

Information on the integrals obtained by the sector decomposition procedure, and the asymptotics of the integral will be given in Sect. 2. Section 3 describes the proposed numerical methods, including a combination of rank-1 lattice QMC integration and extrapolation. Section 4 presents numerical results for the fourth-order coefficients in the expansion of the illustrated three-loop two-point integrals, obtained using a GPU accelerator. Finally, the conclusions are presented in Sect. 5.

2 Formulation

We consider a Laurent expansion of the integral in Eq. (1),

$$I = \frac{1}{\varepsilon^2}C_{-2} + \frac{1}{\varepsilon}C_{-1} + C_0 + \varepsilon C_1 + \cdots\,. \tag{6}$$

The first three terms, i.e., the UV divergent coefficients, are required for the renormalization. As the genuine difference between BB and RS arises from the term $y_1y_2y_3y_4$ in W, it follows from the sector decomposition procedure (see, e.g., [1–3,13,14,16]) described below that the UV divergent coefficients are common

to both. The coefficients $C_k(k=-2,-1,0)$ of these diagrams can be expressed by the following formulae:

$$C_{-2}=1, \quad C_{-1}=2-3j_s^{(1)}(s),$$
$$C_0=8\{c_0+c_1j_s^{(1)}(s)+c_2j_s^{(2)}(s)-3(J_1(s)+J_2(s)+J_3(s))\} \tag{7}$$

where $c_0=-0.210710$, $c_1=(3/4)(\log 2-1)$, $c_2=9/16$,

$$j_s^{(k)}(s)=\int_0^1 dz\,(\log P_0(s))^k, \tag{8}$$

$$J_{1,2}(s)=\frac{1}{2}\int_0^1 dz\int_0^1 dt\int_0^1 du\,(u,1)\frac{\log P_{1,2}(s)}{(1+u+tu)^2}, \tag{9}$$

$$J_3(s)=\frac{1}{2}\int_0^1 dz\int_0^1 dt\int_0^1 du\,\frac{1}{t}\left(\frac{\log P_3(s)}{(1+u+tu)^2}-\frac{\log P_0(s)}{(1+u)^2}\right), \tag{10}$$

$$P_0(s)=m^2-sz\bar{z}, \quad P_{1,2,3}(s)=m^2(1+t+tu)-s\left((tu,t,1)z\bar{z}+\frac{tuz^2}{1+u+tu}\right). \tag{11}$$

The s-dependent functions have a threshold. While $j_s^{(1)}$, $j_s^{(2)}$ and J_3 have the normal threshold at $s=4m^2$, that of J_1 and J_2 is unphysical: The threshold of J_1 is at $s=8m^2$, where $z=3/4$, $t=1$, $u=1$ and that of J_2 is at $s=2(2+\sqrt{3})m^2$ where $z=\sqrt{3}/(\sqrt{3}+1)$, $t=1$, $u=(\sqrt{3}-1)/2$.

The coefficient C_1 is difficult to determine and we calculate C_1 numerically. We confirmed that analytical coefficients $C_k(k=-2,-1,0)$ in Eq. (7) coincide with the results obtained by the numerical calculations.

2.1 Sector Decomposition

The sector decomposition procedure is implemented as follows. We label sectors by $\alpha=(k,l,m,n,p)$, permutation of $(1,2,3,4,5)$. In each sector, we perform the variable transformation.

$$y_k=s_k, \quad y_l=s_kt_l, \quad y_m=s_kt_lu_m, \quad y_n=s_kt_lu_mv_n, \quad y_p=s_kt_lu_mv_nw_p. \tag{12}$$

Though there are $5!=120$ sectors, we only consider 15 sectors due to the symmetry as we study the equal mass case.

The final result is given by

$$I=8\sum_\alpha I(\alpha) \tag{13}$$

where α runs over the 15 permutations:

$$12345, 12354, 12534, 13245, 13254,$$
$$13425, 13452, 13524, 13542, 15234,$$
$$15324, 15342, 51234, 51324, 51342.$$

The U function is a sum of monomials, which are the product of three different y's. While there can be 10 such monomials of $y_1, \cdots, y_5$, U consists of 8 of them. If a $y_k y_l y_m$ is absent in U, i.e., k, l, m is the permutation of $(1, 2, 5)$ or $(3, 4, 5)$, we call the sectors/integrals of Type-II. The rest of sectors/integrals are Type-I.

By this transformation, the U- and W-functions in each sector are expressed as

$$\begin{array}{ll} U = s_k^3 t_l^2 u_m f & W = s_k^4 t_l^2 u_m q \quad (\text{Type} - \text{I}) \\ U = s_k^3 t_l^2 u_m v_n f & W = s_k^4 t_l^2 u_m v_n q \quad (\text{Type} - \text{II}) \end{array} \tag{14}$$

$$\frac{V}{s_k} = G = m^2(1 + t_l(1 + u_m(1 + v_n(1 + w_p)))) - s\frac{q}{f}. \tag{15}$$

When it is redundant, the suffix of s, t, u, v, w is omitted. Thus U and W are replaced by f and q. For instance, when $\alpha = 12345$, $f = 1 + v + uv + uvw + uv^2w + tuv + tuvw + tuv^2w$ and $q = tuv(1 + wz + vwz + uvw\bar{z} + uvw^2\bar{z}z + uv^2w^2\bar{z}z + tuvw\bar{z} + tuvw^2\bar{z}z + tuv^2w^2\bar{z}z)$. Please note that s_k appears in neither f nor q and that f is a regular function.

The integral is given by the following formula. Functions f and q are defined in each sector α.

$$(\text{Type} - \text{I}) \qquad I(\alpha) = \int_0^1 dz \int_0^1 dt \int_0^1 du \int_0^1 dv \int_0^1 dw \frac{hv}{t^{1-2\varepsilon} u^{-\varepsilon}} H, \tag{16}$$

$$(\text{Type} - \text{II}) \qquad I(\alpha) = \int_0^1 dz \int_0^1 dt \int_0^1 du \int_0^1 dv \int_0^1 dw \frac{h}{t^{1-2\varepsilon} u^{-\varepsilon} v^{1-\varepsilon}} H, \tag{17}$$

where $h = y_5/s_k$ and $H = G^{-3\varepsilon}/f^{2-\varepsilon}$.

2.2 Integrands

The sector decomposition procedure allows us to easily expand the integrands of the Type-I and Type-II integrals in ε. In this section, we present the resulting integrands, F, for the 15 sectors. Each F corresponds to the integrand of either the integral in Eq. (16) or Eq. (17). These expressions can be used to evaluate each term analytically and/or numerically. We denote $f_0 = f|_{t=0,v=0}$, $f_a = f|_{t=0}$ and $f_b = f|_{v=0}$ and use a similar notation for G.

$I(12345)$, $I(12354)$, $I(13245)$, $I(13254)$, $I(13425)$, $I(13452)$, $I(13524)$, $I(13542)$, $I(15324)$ and $I(15342)$. The integrals are of Type-I.

$$F = \frac{hv}{tf^2} + \frac{hv}{tf^2}(\log f + 2\log t + \log u - 3\log G)\varepsilon + O(\varepsilon^2) \tag{18}$$

and the s-dependent part arises from the G function.

***I*(12534) and *I*(15234).** The integrals are of Type-II and the integrand can further be written as $F = F_0 + F_1$, with

$$F_0 = \frac{h}{tf_b^2}\frac{1}{\varepsilon} + \frac{h}{tf_b^2}\left(\log f_b + 2\log t + \log u - 3\log G_b\right) + \frac{h}{2tf_b^2}(\log f_b + 2\log t + \log u - 3\log G_b)^2\varepsilon + O(\varepsilon^2) \tag{19}$$

and the s-dependent part comes from the G_b function;

$$F_1 = \frac{h}{tv}\left(\frac{1}{f^2} - \frac{1}{f_b^2}\right) + \frac{h}{tv}\left(\frac{\log f + 2\log t + \log u + \log v - 3\log G}{f^2} - \frac{\log f_b + 2\log t + \log u + \log v - 3\log G_b}{f_b^2}\right)\varepsilon + O(\varepsilon^2) \tag{20}$$

and the s-dependent part appears from the G and G_b functions.

***I*(51234).** The integral is of Type-II and the integrand can be written as $F = F_{00} + F_{01} + F_{10} + F_{11}$, where F_{01} vanishes completely and is therefore omitted.

$$F_{00} = \frac{h}{2f_0^2}\frac{1}{\varepsilon^2} + \frac{h}{2f_0^2}\left(\log f_0 + \log u - 3\log G_0\right)\frac{1}{\varepsilon} + \frac{h}{4f_0^2}\left(\log f_0 + \log u - 3\log G_0\right)^2 + \frac{h}{12f_0^2}\left(\log f_0 + \log u - 3\log G_0\right)^3\varepsilon + O(\varepsilon^2) \tag{21}$$

and the s-dependent part comes from the G_0 function;

$$F_{10} = \frac{h}{t}\left(\frac{1}{f_b^2} - \frac{1}{f_0^2}\right)\frac{1}{\varepsilon} + \frac{h}{t}\left(-\frac{\log f_0 + 2\log t + \log u - 3\log G_0}{f_0^2} + \frac{\log f_b + 2\log t + \log u - 3\log G_b}{f_b^2}\right) + \frac{h}{2t}\left(-\frac{(\log f_0 + 2\log t + \log u - 3\log G_0)^2}{f_0^2} + \frac{(\log f_b + 2\log t + \log u - 3\log G_b)^2}{f_b^2}\right)\varepsilon + O(\varepsilon^2) \tag{22}$$

and the s-dependent part arises from the G_0 and G_b functions;

$$F_{11} = \frac{h}{tv}\left(\frac{1}{f^2} - \frac{1}{f_b^2}\right) + \frac{h}{tv}\left(-\frac{\log f_b + 2\log t + \log u + \log v - 3\log G_b}{f_b^2} + \frac{\log f + 2\log t + \log u + \log v - 3\log G}{f^2}\right)\varepsilon + O(\varepsilon^2) \tag{23}$$

and the s-dependent part comes from the G and G_b functions.

***I*(51324) and *I*(51342).** The integrals are of Type-I and the integrand can be written as $F = F_0 + F_1$, where

$$F_0 = \frac{hv}{2f_a^2}\frac{1}{\varepsilon} + \frac{hv}{2f_a^2}\left(\log f_a + \log u - 3\log G_a\right) + \frac{hv}{4f_a^2}\left(\log f_a + \log u - 3\log G_a\right)^2\varepsilon + O(\varepsilon^2) \tag{24}$$

and the s-dependent part appears from the G_a function;

$$F_1 = \frac{hv}{t}\left(\frac{1}{f^2} - \frac{1}{f_a^2}\right) + \frac{hv}{t}\left(-\frac{\log f_a + 2\log t + \log u - 3\log G_a}{f_a^2} + \frac{\log f + 2\log t + \log u - 3\log G}{f^2}\right)\varepsilon + O(\varepsilon^2) \tag{25}$$

and the s-dependent part comes from the G and G_a functions.

2.3 Asymptotics

In the numerical evaluation, we consider the integrals for the coefficient C_1 as a function of ϱ and construct a sequence of approximations using an appropriate numerical integration method with a decreasing sequence of ϱ. Then we perform an extrapolation to the limit as $\varrho \to 0$. This method for an extrapolation of the sequence relies on the existence of an asymptotic expansion of each integral $\mathcal{I}(\varrho) \sim \mathcal{I} + a_1\phi_1(\varrho) + a_2\phi_2(\varrho) + \cdots$, as $\varrho \to 0$. Although linear or nonlinear extrapolation methods can be explored under certain conditions on the functions $\phi_\ell(\varrho)$, in this paper we use a nonlinear method. The details of the numerical integration and the extrapolation method used in this paper are described in the next section.

3 Numerical Methods

3.1 Lattice QMC Integration

The QMC integration used in this work is based on lattice rules. Rank-1 lattice rules provide an integral approximation over the unit cube of the form

$$\mathcal{Q}f = \mathcal{Q}(\mathbf{z}, N)f = \frac{1}{N}\sum_{j=0}^{N-1} f(\{\frac{j}{N}\mathbf{z}\}) \tag{26}$$

Here, N, the sample size, is a large prime, $\{\mathbf{x}\}$ denotes the vector obtained by taking the fractional part of each component of $\mathbf{x}$ and $\mathbf{z}$ is a generator vector of integer components determined by the specific algorithm for a given N (see [24]). In the current work, N is generally between $\mathcal{O}(10^6)$ and $\mathcal{O}(10^8)$.

A sequence embedding of a lattice rule is defined in [24], based on a rank-1 rule Q_0 and where each rule of the sequence applies a scaled copy of Q_0 to the subregions obtained by subdividing $\mathcal{K}_d$ into m equal parts in each of r coordinate directions. Denoting

$$\mathcal{Q}_r f = \mathcal{Q}_r^{(m)}(\mathbf{z}, N)f = \frac{1}{m^r N}\sum_{k_r=0}^{m-1}\cdots\sum_{k_1=0}^{m-1}\sum_{j=0}^{N-1} f\left(\{\frac{j}{N}\mathbf{z} + \frac{1}{m}(k_1,\ldots,k_r,0,\ldots,0)\}\right) \tag{27}$$

for N and m relatively prime, the points of Q_r are embedded in Q_{r+1} for $0 \le r < d$. Q_r has $m^r N$ points and is of rank r for $1 \le r \le d$, and Q_d is composed of m^d scaled copies of Q_0. An error estimate for Q_d is calculated based on a sequence of rules $Q^{(i)}$, $1 \le i \le d$, of order $m^{d-1}N$ and embedded in Q_d [24]. The results for the 5-dimensional integrals ($d = 5$) in Sect. 4 of this paper are obtained with $m = 4$.

3.2 Transformation

Since lattice rules are applied with a periodizing transformation of the integrand, we select a transformation that also suppresses singularities on the boundaries of the domain. We make use of Sidi's Ψ_m-transformation [21–23], $\Psi_m(t) = \frac{\int_0^t (\sin\pi u)^m}{\int_0^1 (\sin\pi u)^m}$. In this paper we use the Ψ_6-transformation and it is given as

$$\Psi_6(t) = t - (45\sin(2\pi t) - 9\sin(4\pi t) + \sin(6\pi t))/(60\pi), \quad \Psi_6'(t) = \frac{16}{5}\sin^6(\pi t). \tag{28}$$

3.3 Extrapolation

With the goal of obtaining the extrapolated results numerically, we use the ϵ-algorithm [20,27] with an implementation based on the code of *qext()* from the QuadPack package [19]. The computation produces a triangular table, with a

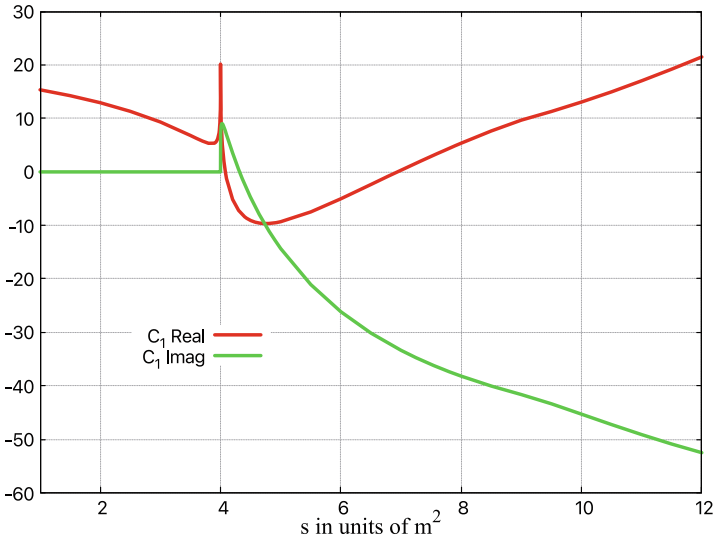


Fig. 2. Baseball C_1 by QMC integration

sequence of integral approximations as the entry sequence in the first column. Each subsequent column is generated from neighboring entries of the previous ones. The *qext()* procedure computes a new lower diagonal of the table when a new entry is available, together with a measure of the distance of each element from preceding neighboring elements. In the construction of the triangular table, a table element *"new"* relative to preceding $e_0,\ e_1,\ e_2,\ e_3$ is obtained as

$$\begin{matrix} & e_0 & \\ e_3 & e_1 & new \\ & e_2 & \end{matrix} = e_1 + 1/(1/(e_1 - e_3) + 1/(e_2 - e_1) - 1/(e_1 - e_0)) \tag{29}$$

and the distance for *new* is set to

$$distance = |e_2 - e_1| + |e_1 - e_0| + |e_2 - new|. \tag{30}$$

Then *qext()* returns the new lower diagonal element with the smallest value of the distance estimate as the extrapolation result.

4 Results

For the numerical results, we focus on the computation of the coefficient C_1 of the linear term in a Laurent expansion. We studied the expansions and their lower-order coefficients in earlier work, e.g., [7,9] for the Baseball and various other diagrams. Figure 2 shows plots of Baseball (BB) C_1 as a function of the squared momentum s and Fig. 3 shows a similar plot for Running Shirt (RS).

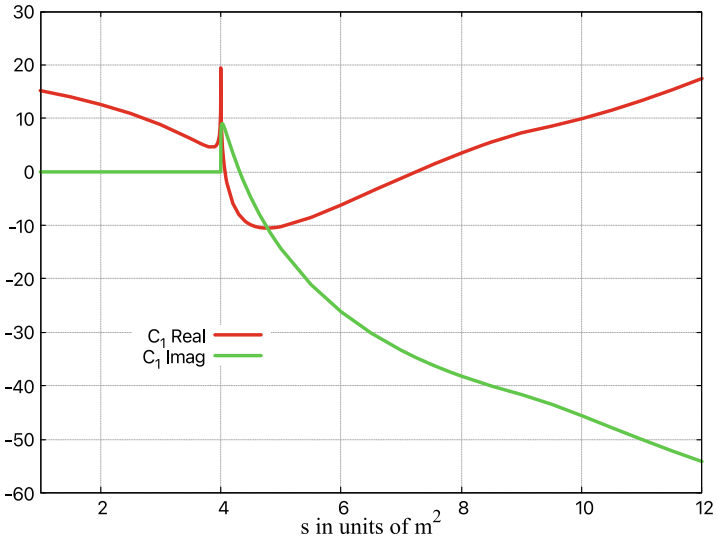


Fig. 3. Running Shirt C_1 by QMC integration

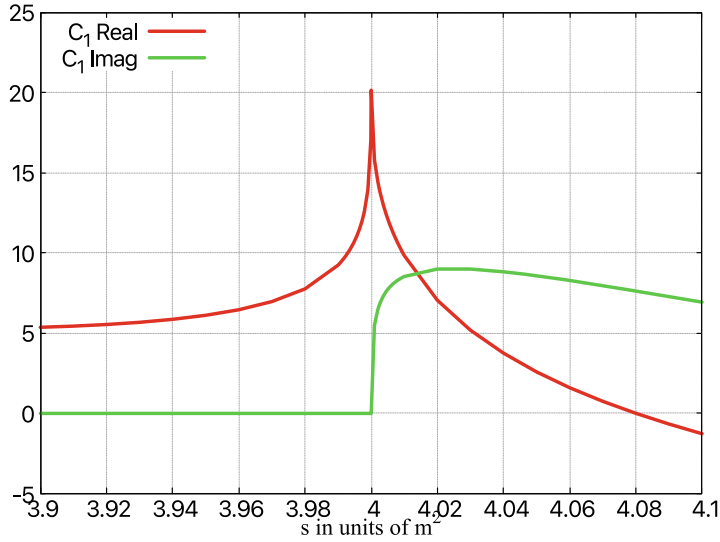


Fig. 4. Baseball C_1 around s = 4.0

As the diagrams have a threshold (singularity) at $s = 4$, we plot the behavior around $s = 4$ for BB and RS in Figs. 4 and 5, respectively. Note that the real part peaks at the threshold and the imaginary part emerges (for $s \geq 4$).

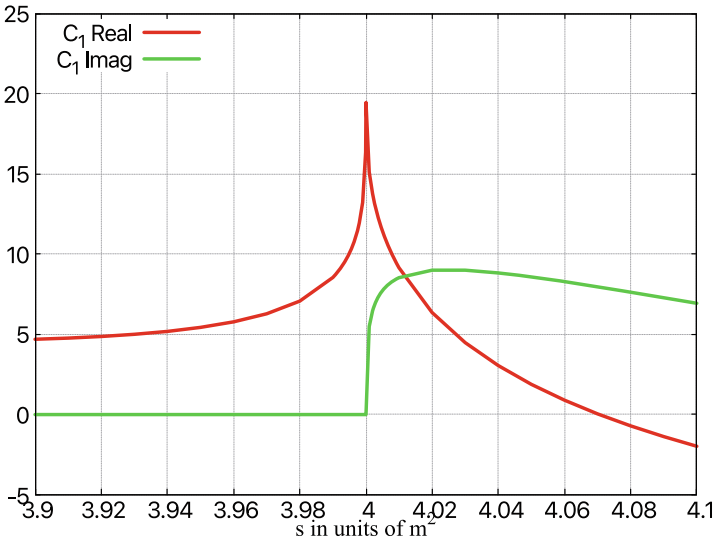


Fig. 5. Running Shirt C_1 around s = 4.0

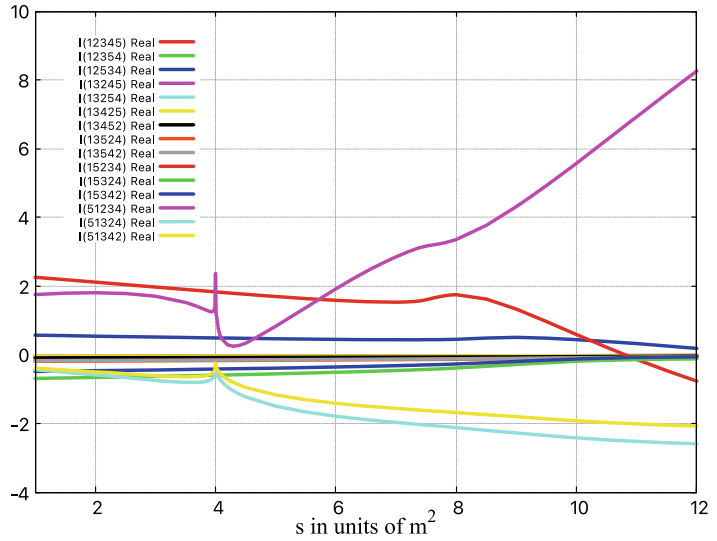


Fig. 6. Baseball C_1 real part 15 sectors by QMC integration

In the numerical computation, we set all masses equal, $m_r = 1, (r = 1, \cdots, 6.)$ For the QMC integration, we use the parameters $N = 400\text{M}$ [5] and $m = 4$, and for the extrapolation we take $\varrho = 1.2^{-24}, \cdots, 1.2^{-38}$ for most values of s. The numerical results are in agreement with those obtained by pySecDec [4], with a

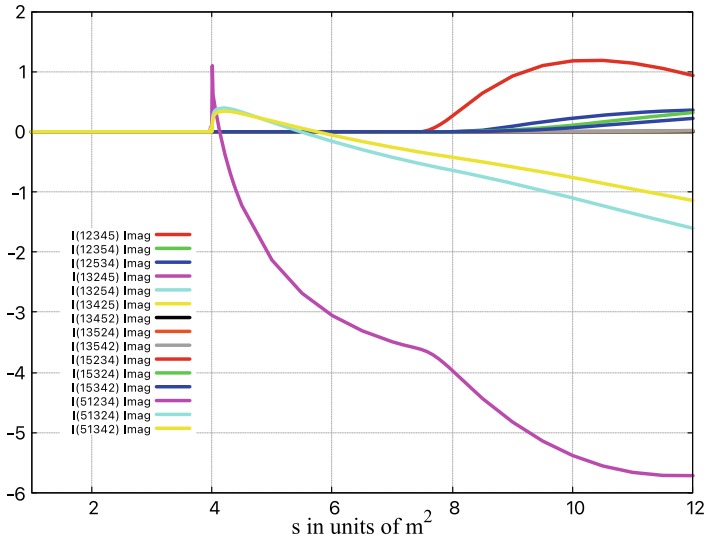


Fig. 7. Baseball C_1 imag part 15 sectors by QMC integration

precision of about five to six digits. For example, for $s = 1$ the coefficient C_1 of the real part of BB is 15.351339, while the corresponding result from pySecDec is 15.351340 $\pm 3.8 \times 10^{-7}$, and C_1 of the real part of RS is 15.200684 and from pySecDec it is 15.200685 $\pm 3.9 \times 10^{-7}$.

Figures 6 and 7 give plots of BB C_1 (real and imaginary) for the 15 sectors. The 15 solid curves correspond to the contributions from individual sectors. The sector decomposition method provides an effective means of isolating UV divergences and improving the convergence of individual integrands. However, because the integration domain is partitioned, the energy dependence of each component may exhibit unphysical behavior associated with artificial thresholds. As illustrated in Fig. 2, once the contributions from all sectors are summed, the final result shows a natural and physically consistent behavior.

The QMC integrations were performed using an NVIDIA GH200 Grace Hopper Superchip [18] equipped on the Miyabi supercomputer system [12]. The extrapolations are performed using Intel(R) Xeon(R) Gold 6230 CPU @ 2.10 GHz. The integrations and extrapolations were computed in double precision.

The elapsed time for the QMC integration in each s value with $m = 4$ ($400M \times 4^5$ sample points) was around 11900 s using a GH200 and that for the extrapolation was negligible. The elapsed time for our method depends on the total number of sampling points. In contrast, the elapsed time for pySecDec depends on the value of s. For values of s in the range $1.0 < s < 12.0$, the observed elapsed time ranged from 36 s to 3026 s on an NVIDIA GV100. Further

improvements to reduce the elapsed time of our method will be pursued by refining the CUDA code and implementing MPI parallelization.

5 Conclusions

In this paper, we studied two three-loop two-point Feynman diagrams. In particular, we evaluated the fourth term in the Laurent expansion for ε, where the difference between the two diagrams appears.

The Feynman integral becomes singular above the threshold, and some other methods use a deformation of the contour to avoid the zero of V, while the control of the deformation is sometimes complicated and tedious. In our numerical method, the zero of V is avoided by taking the regularization parameter ρ as a finite number. When one can use a good precision integration and extrapolation, it is possible to obtain the value of the integral, though the cost of computation increases for the evaluation of the sequence of approximations.

We obtained the energy dependence of the loop integrals. The methods employed in this work, the combination of the GPU-accelerated QMC integration and the numerical extrapolation, have been shown to be effective. We also obtained results for values of s beyond the threshold. The numerical results achieve an accuracy of five to six digits, which is sufficient for phenomenological applications. As a future direction, the present approach can be extended to other three-loop diagrams and higher order loop integrals.

Acknowledgements. We acknowledge the support by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant Numbers 25K07331. This research used computational resources of Miyabi-G provided by the Multidisciplinary Cooperative Research Program in the Center for Computational Sciences, University of Tsukuba.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Binoth, T., Heinrich, G.: An automized algorithm to compute infrared divergent multi-loop integrals. Nucl. Phys. B **585**, 741–759 (2000). arXiv:hep-ph/0004013
2. Binoth, T., Heinrich, G.: Numerical evaluation of multi-loop integrals by sector decomposition. Nucl. Phys. B **680**, 375 (2004). arXiv:hep-ph/0305234
3. Binoth, T., Heinrich, G.: Numerical evaluation of phase space integrals by sector decomposition. Nucl. Phys. B **693**, 134 (2004). arXiv:hep-ph/0402265
4. Borowka, S., Heinrich, G., Jahn, S., Jones, S.P., Kerner, M., Schlenk, J.: A GPU compatible quasi-Monte Carlo integrator interfaced to pySecDec. Comput. Phys. Commun. **240**, 120–137 (2019). arXiv:1811.11720
5. de Doncker, E., Almulihi, A., Yuasa, F.: Transformed lattice rules for Feynman loop integrals. IOP Ser. J. Phys. Conf. Ser. (JPCS) **1136**(012002) (2018). https://doi.org/10.1088/1742-6596/1136/1/012002

6. de Doncker, E., Ishikawa, T., Kato, K., Yuasa, F.: Analytic and numerical approaches for depictive 3-loop integrals using sector decomposition. Prog. Theor. Exp. Phys. (2024). arXiv:2405.13286v1 [hep-ph]. https://doi.org/10.48550/arXiv.2405.13286
7. de Doncker, E., Yuasa, F., Ishikawa, T., Kato, K.: 3-loop Feynman integral extrapolations for the baseball diagram. In: The 22nd International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2024) (2024 To appear)
8. de Doncker, E., Yuasa, F.: Self-energy Feynman diagrams with four loops and 11 internal lines. In: Lecture Notes in Computer Science (LNCS), vol. 12953, pp. 160–175. Springer (2021)
9. de Doncker, E., Yuasa, F., Ishikawa, T., Kato, K.: Methods for extrapolation of 3-loop 2-point Feynman integrals. In: Lecture Notes in Computer Science (LNCS), vol. 15890, pp. 20–30. Springer (2025). https://doi.org/10.1007/978-3-031-97606-3_2
10. de Doncker, E., Yuasa, F., Kato, K., Ishikawa, T., Kapenga, J., Olagbemi, O.: Regularization with numerical extrapolation for finite and UV-divergent multi-loop integrals. Comput. Phys. Commun. **224**, 164–185 (2018). arXiv:1702.04904
11. de Doncker, E., Yuasa, F., Olagbemi, O., Ishikawa, T.: Large scale automatic computations for Feynman diagrams with up to five loops. In: Lecture Notes in Computer science (LNCS), vol. 12253, pp. 145–162. Springer (2020)
12. Hanawa, T., et al.: Miyabi: new supercomputer system of JCAHPC (2024). https://doi.org/10.24669/axies.2024.0_154. (in Japanese)
13. Heinrich, G.: Sector decomposition. Int. J. Mod. Phys. A **23**, 1457 (2008). arXiv:0803.4177
14. Heinrich, G.: Collider physics at the precision frontier. Phys. Rep. **922**, 1 (2021). arXiv:2009.00516
15. Laporta, S.: High-precision calculation of multi-loop Feynman integrals by difference equations. Int. J. Mod. Phys. A **15**, 5087–5159 (2000). arXiv:hep-ph/0102033
16. Nakanishi, N.: Graph Theory and Feynman Integrals. Gordon and Breach, New York (1971)
17. Nogueira, P.: Automatic Feynman graph. J. Comput. Phys. **105**, 279–289 (1993). https://doi.org/10.1006/jcph.1993.1074
18. NVIDIA. https://www.nvidia.com/en-us/data-center/grace-hopper-superchip/
19. Piessens, R., de Doncker, E., Überhuber, C.W., Kahaner, D.K.: QUADPACK, A Subroutine Package for Automatic Integration. Springer Series in Computational Mathematics, vol. 1. Springer-Verlag (1983)
20. Shanks, D.: Non-linear transformations of divergent and slowly convergent sequences. J. Math. Phys. **34**, 1–42 (1955)
21. Sidi, A.: A new variable transformation for numerical integration. Int. Ser. Numer. Math. **112**, 359–373 (1993)
22. Sidi, A.: Extension of a class of periodizing transformations for numerical integration. Math. Comp. **75**(253), 327–343 (2005)
23. Sidi, A.: Further extension of a class of periodizing variable transformations for numerical integration. J. Comput. Appl. Math. **221**, 132–149 (2008)
24. Sloan, I., Joe, S.: Lattice Methods for Multiple Integration. Oxford University Press (1994)
25. Smirnov, A.V.: FIESTA4: optimized Feynman integral calculations with GPU support. Comput. Phys. Commun. **204**, 189–199 (2016). https://doi.org/10.1016/j.cpc.2016.03.013. arXiv:1511.03614

26. Usovitsch, J., Dubovyk, I., Riemann, T.: MBnumerics: numerical integration of Mellin-Barnes integrals in physical regions. PoS LL **046**, 1810.04580 (2018). [hep-ph]. https://doi.org/10.22323/1.303.0046
27. Wynn, P.: On a device for computing the $e_m(s_n)$ transformation. Math. Tables Aids Comput. **10**, 91–96 (1956)

Development and Validation of the Computational Code GEM for Simulating Quantum Few-Body Systems through Optimization on the Supercomputer Fugaku

Shigeyoshi Aoyama[1,5(✉)], Yusuke Tsunoda[2,5], Noritaka Shimizu[2,5], Hideo Matsufuru[1,5], Issaku Kanamori[3,5], and Emiko Hiyama[4,5]

[1] Computing Research Center, High Energy Accelerator Research Organization (KEK), Tsukuba, Ibaraki 305-0801, Japan
shigeyoshi.aoyama@kek.jp
[2] Center for Computational Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba 305-8577, Japan
[3] RIKEN Center for Computational Science, Kobe 650-0047, Japan
[4] Department of Physics, Tohoku University, Sendai 980-8578, Japan
[5] RIKEN Nishina Center, Wako 351-0198, Japan

Abstract. The Gaussian expansion method (GEM) is a prominent approach widely used for precision calculations in few-body systems. While the calculation of its matrix elements is considerably complex, to alleviate these computational complexities, the Gauss-Lobe technique was proposed by one of the authors (Hiyama). This approach has been implemented in a well-established computational code (GEM code) that has been applied to the analysis of numerous few-body systems over the years. Furthermore, it is renowned for its high precision and versatility in few-body system calculations. However, the core of this code was developed more than thirty years ago and, while it has undergone numerous updates over the decades, it is not architected to fully exploit the performance of modern massively parallel computers, such as the Supercomputer Fugaku as well as recent GPU-based systems.

In this paper, we report on and discuss the modernization of GEM code, which has long been based on a design for single-processor systems. We present its implementation for high-performance hybrid parallel computing on the Supercomputer Fugaku and evaluate its computational performance. We demonstrate its performance using the ^{4}He pentamer, a five-body system for which detailed analysis is currently highly anticipated.

Through the modernization of the GEM code in this study, the hybrid parallelization achieved a 122.5-fold speed-up using 120 nodes, demonstrating a so-called "super-linear speed-up" that exceeds the ideal value. This result clearly highlights the significant utility and effectiveness of the parallelization strategy implemented in the modern GEM code. Furthermore, this paper describes a practical technique for adjusting the ratio

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 492–507, 2027.
https://doi.org/10.1007/978-3-032-30521-3_31

between MPI parallelism and thread parallelism without modifying the original GEM source code. This technique enables efficient parallelization for six-body or even larger few-body systems.

Keywords: few-body system · massive parallel computing · HPC

1 Introduction

Numerical approaches to quantum few-body systems have emerged as a pivotal methodology in the analysis of quantum systems. Given the recent breakthroughs in generative AI and machine learning, there is growing anticipation for realizing quantum computing to accelerate these technologies. Consequently, the precise simulation of quantum few-body systems not only represents a challenge to fundamental issues at the frontiers of conventional science but is also expected to serve as a vital theoretical foundation for the practical realization of quantum computing.

Furthermore, many intriguing yet unresolved issues remain in the field of few-body physics, such as the excited states of the ^{4}He pentamer (five-body system) investigated in this paper. The ^{4}He trimer (three-body system) and ^{4}He tetramer (four-body system) are characterized by a compact ground state located near the van der Waals potential pocket of several angstroms; on the other hand, their first excited states are S-wave halo-like states that extend to several hundred angstroms, far beyond the potential pocket. Such states are also considered to be related to Efimov states [1] and Bose—Einstein condensation [2–4], which continue to be of profound physical interest. Similarly, systems of five or more ^{4}He atoms are expected to exhibit such a state with an S-wave halo tail extending to several hundred angstroms near the ^{4}He one-atom separation threshold.

To solve quantum few-body systems, various sophisticated methods have been developed, including the Gaussian Expansion Method (GEM) [5–7], an extension of the Faddeev theory to N-body systems [8,9], Hyperspherical Harmonics (HH) [10], the No-Core Shell Model (NCSM) [11,12], and Green's Function Monte Carlo (GFMC) [13], etc. Among these, GEM is widely recognized for its exceptional capability to rigorously solve Hamiltonians in which complex parti- cle rearrangements occur, and where ground and excited states must be treated simultaneously despite their vastly different spatial scales. This makes GEM particularly effective for systems such as those addressed in this study, where a compact ground state and a significantly extended halo-like excited state must be treated within a single unified framework. While various equivalent approaches to the Gaussian Expansion Method (GEM) have been developed by numerous research groups [14–18] due to its high utility, this paper employs the Gauss-Lobe basis framework. In this framework, as proposed by one of the present authors (Hiyama), the angular momentum integration is formulated as a limit of Gaussian integrations, making the hand-calculation of complex few-body integrals much more manageable.

Beyond the ^{4}He pentamer addressed in this study, various five-body systems remain at the forefront of contemporary few-body physics. Notable examples include nuclear systems such as ^{8}He (^{4}He+n+n+n+n) and ^{5}H (p+n+n+n+n), as well as hypernuclei like ${}^{5}_{\Lambda}$H ($p+n+n+n+\Lambda$). Furthermore, kaonic nuclei such as $K^- + p + n + n + n$ and quark cluster systems like the Θ^+ pentaquark ($uudds$) continue to be subjects of intense theoretical and experimental interest.

Furthermore, by increasing the number of particles in few-body calculations to ten or more and comparing the results with those of many-body systems, it is expected that the fundamental principles underlying many-body physics can be more firmly established. From this perspective, the authors have been making efforts to increase the number of particles that can be treated within the GEM framework. Figure 1 schematically illustrates the ongoing and future expansions of our methodology. The blue text represents the entire project, while the red (this work) and blue texts indicate the specific approaches taken toward it.

At CSHPC 2025, some of authors (S. Aoyama and T. Myo) reported an extension of the stochastic variational method (SVM) [21–23], referred to as SVM$^+$, as a machine learning-inspired approach [24] (the upper part of Fig. 1), specifically aimed at reducing memory consumption when increasing the number of particles. As a benchmark, we have already achieved an approximately 99% reduction in memory usage at the three-body system stage. Similarly, the primary bottleneck when increasing the number of particles is the computational cost of calculating matrix elements. In this paper, we report on the development of a newly optimized GEM code focused on this challenge and the resulting performance improvements.

At HPC Asia 2026, some of the present authors (E. Hiyama, I. Kanamori, and H. Matsufuru) reported on solving large-scale eigenvalue problems in quantum few-body systems (as shown in the left part of Fig. 1) and specialized numerical algorithms of diagonalization (as shown in the right part of Fig. 1) [25]. In the Gaussian Expansion Method (GEM), the strong linear dependence of the basis functions results in dense matrices with large off-diagonal elements. In such cases, it is well known that even minor numerical inaccuracies can be amplified during the diagonalization of large-dimensional matrices, making it difficult to obtain high- precision solutions. Consequently, solving the eigenvalue problem itself becomes a challenging task. In Ref. [25], a systematic analysis of the numerical errors occurring during actual diagonalization was performed using EigenExa [25], an MPI-parallelized eigensolver had developed on the supercomputer Fugaku.

The original Gaussian Expansion Method (GEM) code was developed over 30 years ago for three-body systems and has since been extended and applied to four- and five-body systems. Having been utilized for decades, this code is not only remarkably stable but also capable of maintaining a high precision of over ten decimal places, making it a highly versatile tool applicable to a wide range of physical systems.

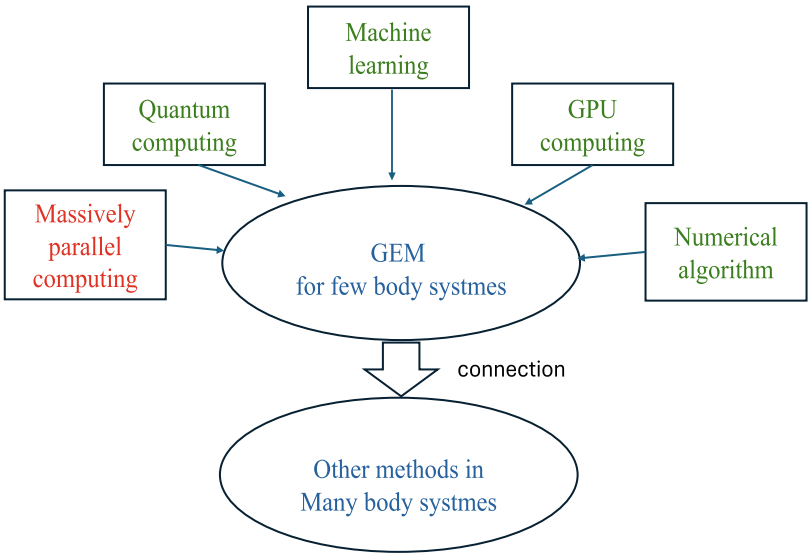


Fig. 1. Schematic diagram of GEM's development and expansion.

However, this legacy code was not designed for modern hybrid parallelization or GPU acceleration. Furthermore, the original GEM code does not account for the efficient utilization of memory and registers, such as SIMD (Single Instruction, Multiple Data) operations. Consequently, the code typically achieves only a few percent, specifically around 3–4% (in present computation), of the peak FLOPS specifications, and its memory bandwidth utilization remains below 1%. Thus, there remains substantial potential for further acceleration—on the order of 10^7 at least—through code-level optimizations (such as SIMD), intra-node GPU parallelization, and large-scale MPI parallelization ($10\times1{,}000\times1{,}000$). Such advancements could bring currently inaccessible six-, seven-, and eight-body systems within reach of precision calculations.

Although the supercomputer Fugaku, Japan's current flagship supercomputer, is not equipped with GPUs, its massive scale of over 150,000 nodes. If approximately 10% of the total node count is utilized, there is the potential for tens of thousands of times faster computation without GPUs, provided that the parallel efficiency relative to the node count can be brought close to 100%. Therefore, in this paper, we aim as a first step to maximize this speed-up ratio during the parallelization of the GEM code (the left part of Fig. 1). Furthermore, optimization of memory bandwidth and register utilization, which are potential avenues for further acceleration on Fugaku, is beyond the scope of this study and will be addressed in future work, such as during the extension of the code to six-body systems or larger.

The remainder of this paper is organized as follows. In Sect. 2, the framework relating GEM used in the present calculation is briefly explained. In Sect. 3,

we present the calculation results and provide a detailed discussion. Finally, a summary and our conclusions are given.

2 Method

2.1 Gaussian Basis Function

Here, we briefly explain GEM for a five-body bosonic system with Jacobi coordinate. The detail of GEM is given in review papers [7,26]. In GEM, the wave function Ψ_{JM} (J is the total angular momentum and M is its z-component) of the Schrödinger equation,

$$(H - E)\Psi_{JM} = 0, \tag{1}$$

is obtained by diagonalizing the Hamiltonian in a space spanned by Gaussian basis functions which are constructed on Jacobi coordinates. ^{4}He–^{4}He potential in the Hamiltonian used in this research, $V(r) = V_1 \exp(-\alpha_1 r^2) + V_2 \exp(-\alpha_2 r^2)$, with parameters, $V_1 = 2000\,\mathrm{K}$, $\alpha_1 = 1.0\ \mathrm{Å}^{-2}$, $V_2 = -8.2823\ \mathrm{K}$, $\alpha_2 = 0.082189\ \mathrm{Å}^{-2}$. This two range (attractive and repulsive) Gaussian potential was made to reproduce the observed properties of ^{4}He dimer [27].

In Fig. 2, the Jacobi coordinates for a five-body system is displayed $\boldsymbol{R_i}$ is i-th Jacobi coordinate, L_i is angular momentum, b_i is a width parameter of Gaussian type basis functions. (a) is 4(K)+1 type or in order type which describes the K type for the four-body system plus a ^{4}He atom. (b) is 4(H)+1 type which describes the H type for the four-body system plus a ^{4}He atom. (c) is 3+2 type which describes trimer and dimer which first appear in the five-body system. If a trimer plus dimer like state is important, (c) is more appropriate than (a) or (b).

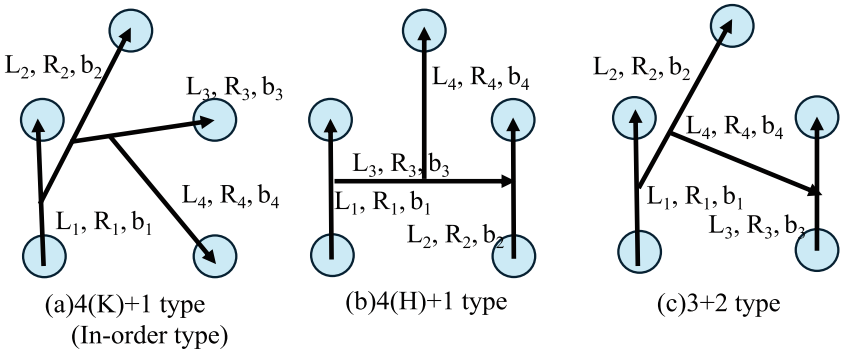


Fig. 2. Jacobi coordinates for five-body system.

The total wave function Ψ_{JM} is written as

$$\Psi_{JM} = \sum_{type=1}^{3} \sum_{i=1}^{i_{\max}} c_i \mathcal{S} \Phi_i^{type}(\boldsymbol{R_1}, \boldsymbol{R_2}, \boldsymbol{R_3}, \boldsymbol{R_4}). \tag{2}$$

where $\mathcal{S}$ is the symmetrization operator for bosonic systems, Φ_i is the five-body basis function, and *type* in Eq. (2) shows the type of Jacobi coordinate as shown in Fig. 2 (type=1: 4(K)+1, type=2: 4(H)+1, type=3: 3+2). In a five-body bosonic system with the same mass, the symmetrization for the wave function is equivalent to rearranging the Jacobi coordinate [7,26]. For example, in a 4(K)+1 type, the number of permutations is $5!/2 = 60$.

The Schrödinger equation is expressed in the matrix-element form as follows:

$$\sum_{j=1}^{N}(H_{ij} - ES_{ij})c_j = 0 \quad (i = 1, ..., N), \tag{3}$$

where the hamiltonian matrix-elements H_{ij} are given by

$$H_{ij} = T_{ij} + V_{ij}. \tag{4}$$

Here, T_{ij}, V_{ij}, and S_{ij} represent the kinetic energy, potential energy, and overlap matrix-elements, respectively. Here, H_{ij} represents the hamiltonian matrix-elements, which consist of the kinetic energy matrix-elements T_{ij} and the potential energy matrix-elements V_{ij}. Additionally, S_{ij} are referred to as the overlap matrix-elements.

The five-body basis function in Eq. (2) is represented as

$$\begin{aligned}\Phi_i(\boldsymbol{R_1}, \boldsymbol{R_2}, \boldsymbol{R_3}, \boldsymbol{R_4}) &= \phi_{n_1 L_1}(R_1)\,\phi_{n_2 L_2}(R_2)\,\phi_{n_3 L_3}(R_3)\,\phi_{n_4 L_4}(R_4) \\ &\times \left[\left[\left[Y_{L_1}(\widehat{R}_1) Y_{L_2}(\widehat{R}_2)\right]_{L_{12}} Y_{L_3}(\widehat{R}_3)\right]_{L_{123}} Y_{L_4}(\widehat{R}_4)\right]_{JM}, \quad (5) \\ i &\equiv \{n_1, n_2, n_3, n_4, L_1, L_2, L_{12}, L_3, L_{123}, L_4\}. \quad (6)\end{aligned}$$

where the R_i is i-th Jacobi coordinate, L_i is the angular momentum of the spherical harmonic function Y_L, and n_i is the basis number. *type* is omitted for simplicity. Radial dependence of the basis functions $\phi_{nL}(R)$ is taken as Gaussians multiplied by R^L with ranges in geometric progression [7,26],

$$\phi_{nL}(R) = N_{nL} R^L \mathrm{e}^{-a_n R^2}, \tag{7}$$

$$a_n = 1/b_n^2, \tag{8}$$

$$b_n = b_{\min}\gamma^{n-1} \quad (n = 1, \ldots, n_{\max}), \tag{9}$$

b_n is a Gaussian width parameter, γ is the common ratio given by $(b_{\max}/b_{\min})^{\frac{1}{n_{\max}-1}}$ and N_{nL} is a normalization constant. The integration of the matrix elements for the angular momentum part is very complicated due to many of the spherical harmonic functions, $Y_{LM}(\hat{R})$, with the permutations of Jacobi coordinates. To avoid this difficulty, we introduce a technique using the following basis function (called Gauss-Lobe):

$$R^L e^{-\nu R^2} Y_{LM}(\hat{R}) = \lim_{\varepsilon \to 0} \frac{1}{(\nu\varepsilon)^L} \sum_{k=1}^{k_{\max}} C_{L,M,k} \exp\left(-\nu\left(\mathbf{R} - \varepsilon \mathbf{D}_{L,M,k}\right)^2\right), \tag{10}$$

where vectors $D_{L,M,k}$ and the order of expansion $k_{\max}$ are determined for L and M. When ε is set to a near zero value, the integration of the angular momentum part can be carried out by a straightforward Gaussian integration [7]. While the left and right sides of Eq. (10) are mathematically equivalent through the limiting procedure, it has been confirmed that the matrix elements calculated using the Gauss-Lobe basis numerically agree with those obtained using the conventional Gaussian basis to high precision.

2.2 Channels

Table 1. Quantum numbers (orbital angular momenta) of the channels, which are considered to be significant for low-energy states of ^{4}He pentamer with $J = 0^+$.

type	No.	L_1	L_2	L_{12}	L_3	L_{123}	L_4
$4(K)+1$	1	0	0	0	0	0	0
	2	0	1	1	1	0	0
	3	0	2	2	2	0	0
	4	2	0	2	2	0	0
	5	2	2	0	0	0	0
$4(H)+1$	6	0	0	0	0	0	0
	7	0	2	2	2	0	0
	8	2	0	2	2	0	0
	9	2	2	0	0	0	0
$3+2$	10	0	0	0	0	0	0
	11	0	2	2	2	0	0
	12	2	0	2	2	0	0
	13	2	2	0	0	0	0

In Table 1, the quantum numbers (orbital angular momenta) of the channels are listed, which are considered to be significant for low-energy states of ^{4}He pentamer with $J = 0^+$. The first column corresponds to the Jacobi coordinates illustrated in Fig. 2. The second column provides channel numbers to make it easy to identify which channels are included in the subsequent calculations. The third and following columns list the orbital angular momenta defined for each set of Jacobi coordinates as described in Eq. (5) and Fig. 2. For example, in the case of the 4(K)+1 type, the total angular momentum J is coupled as $[[[[L_1L_2]_{L_{12}}L_3]_{L_{123}}]L_4]_J$. One of the significant features of the GEM basis is that it is a non-orthogonal basis, where each basis function possesses strong overlaps with others. Furthermore, within each channel, permutations between particles (symmetrization in the case of ^{4}He atoms) are performed. For instance, in the 4(K)+1 type, each state is a superposition of 60 (= 5!/2) basis functions after

boson permutation, which leads to the linear dependence. Consequently, even if Channel 1 and Channel 2 appear at first glance to be orthogonal in their angular momentum and have no overlap, they do, in fact, overlap. If a sufficient number of Gaussian basis functions are superimposed in Channel 1, a large portion of Channel 2's space is effectively covered; thus, Channel 2 serves a complementary role to fill in what remains missing from Channel 1. For this reason, the number of basis functions in Channel 2 can fundamentally be reduced compared to Channel 1 without loss of accuracy. This principle applies to other channels as well; the basis space that would normally be required for a later-added channel is already partially incorporated by the preceding basis functions. Therefore, convergence can be achieved by introducing only a small number of additional basis functions. From a numerical methodology perspective, a key advantage of GEM is the ability to achieve rapid convergence with a relatively small number of basis functions by utilizing this rearrangement of Jacobi coordinates and the strong linear dependence of the Gaussian basis.

While the channels listed in Table 1 do not exhaust all possibilities even from the viewpoint of including subsystem angular momenta up to $L = 2$. For the case of the ^{4}He pentamer in this study, they represent a sufficient selection to demonstrate the practical feasibility of the numerical calculations. This selection provides sufficiently converged results for the objectives of the present paper.

3 Results

3.1 Parallelization

In the development of the present code, we adopted a standard "hybrid parallelization" strategy, utilizing MPI for inter-node parallelism and thread-level parallelism for intra-node operations. The supercomputer Fugaku has relatively small memory capacity per node compared to other major supercomputers. To utilize the limited memory resources efficiently, our code avoids intra-node MPI parallelization. While one might consider a method that saves memory by writing out matrix-elements as they are calculated, such an approach could cause storage I/O delays when scaled to tens of thousands of nodes in the future, placing an excessive load on the shared file system. For this reason, we designed the system to batch writing at the memory capacity limit. Furthermore, as will be shown later, the speed-up ratio for MPI is parallel efficiency of 98.3%. This indicates that no significant delays caused by aggregating data to specific nodes via MPI have occurred at the scale of the five-body system in this study.

While we initially planned to use OpenMP for intra-node parallelization, the matrix-element calculations consist of a simple single-loop structure, which is appropriate for applying the OpenMP to that loop. Consequently, we found that the standard compiler's automatic parallelization provided approximately a 10-fold speed-up. It means that significant speed-up is not expected even if OpenMP is adopted. We mention that a 48-fold speed-up would be achieved using 48 cores on a single node in the ideal case. However, as will be demonstrated later, the current version is expected to reduce the computation time

from approximately one week to just a few tens of minutes, representing less than 1% of the resources on Fugaku. The computational speed achieved with MPI+"automatic parallelization" is sufficient for solving the five-body problem. Therefore, to avoid unnecessary bugs that may arise from manual OpenMP implementation, for this initial version of the five-body system code, we have decided to rely on automatic parallelization for the time being. When computational time and resources become more demanding for extensions to six-body systems or larger, we may revisit the implementation of the remaining potential for the speed-up.

The specific points of parallelization are described below. First, for MPI parallelization, we targeted the double loops over the channels (bra and ket) as the primary parallelization points. For instance, in the case of the 13 channels shown in Table 1, $13 \times (13+1)/2 = 91$ MPI processes can be utilized. However, if the number of channels is small, the degree of parallelism is limited; for a single channel, parallelization is not possible at all.

To circumvent this, we decomposed each channel into several segments using different Gaussian bases. For example, if a single channel is divided into N segments, $N \times (N+1)/2$ parallel processes become possible. We refer to this approach as "virtual channel decomposition". Furthermore, while the calculation time for matrix elements within the same channel is almost identical, it generally varies between different channels. In actual calculations, we improved efficiency by adjusting these division numbers to equalize the computation time across each process, facilitating load balancing.

In GEM, a matrix-element expression contains the Gaussian width parameters of the basis functions for both bra and ket states $(b'_1, b'_2, b'_3, b'_4, b_1, b_2, b_3, b_4)$. For example, if we consider a spatial superposition of 10 basis functions for each of the degrees of freedom, the number of independent matrix elements reaches 10^8. In the actual code, the calculation follows this structure:

```
 do i = 1, 100000000
     ham(i) = matrix(bp1(i), bp2(i), bp3(i), bp4(i),
&                    b1(i), b2(i), b3(i), b4(i))
 enddo
```

While the matrix elements depend on various factors such as the angular momentum specified by each channel, for the sake of simplicity, we have expressed them in the form of a function `matrix()`, highlighting only their dependence on the parameters b. The explicit expressions for the matrix elements in three- and four-body systems have been provided in previous review papers [7,26]. Although we do not show the explicit form of matrix elements for the five-body system, the formulation for the five-body system is a straightforward extension of those cases.

As there are no data dependencies across the iterations of i, the parallelization efficiency for this part was expected to be extremely high. Initially, we planned to implement OpenMP for this loop. However, due to its remarkably simple structure, we found that the compiler's standard auto-parallelization feature alone could achieve approximately a tenfold speed-up.

Here, we describe the "virtual channel decomposition" used to increase the number of MPI processes, as its implementation relies on the loop structure discussed above. In our code, the channel types, angular momenta, and Gaussian range parameters (as shown in Table 1) are provided via an input file to determine the matrix elements to be calculated. For each Jacobi coordinate, these Gaussian range parameters are defined by the initial value $b_i^{(1)}$, the final value $b_i^{(N)}$, and the number of basis functions N_i $(i = 1, 2, 3, 4)$, which consequently determines the common ratio of the geometric progression.

The fundamental idea is to virtually treat a single channel as multiple channels by partitioning these parameters (initial value, final value, and number of bases). For instance, in the case of a $10 \times 10 \times 10 \times 10$ basis set, the last parameter b_4 can be divided into 10 segments, which are then described as 10 distinct virtual channels in the input file. Consequently, the 10^8 iterations per channel mentioned in the previous example are reduced by a factor of 10, to 10^7 iterations per virtual channel.

This approach allows us to shift a portion of the loops typically handled by thread-level parallelism (b_4 and b_4') toward MPI-level parallelism without any modification to the source code itself. While a degree of parallelism in the hundreds is sufficient for intra-node thread parallelism (matching the number of cores per node), the scale required for the full five-body calculation is significantly larger.

For the final calculations of the five-body system, a few hundred nodes are sufficient, for which the current decomposition of (b_4 and b_4') is adequate. However, for massive supercomputers like Fugaku, which features over 150,000 nodes, one can easily extend this logic to the loops corresponding to (b_3 and b_3'). This enables efficient calculations across tens of thousands of nodes by shifting the parallel workload to the MPI level without requiring any code revisions. For systems with six-body or larger, it is anticipated that the required MPI parallelism will reach the scale of thousands or even tens of thousands of processes. In such cases, the technique described above—virtually decomposing channels to balance the workload across the massive MPI resources—becomes increasingly essential.

Since the actual memory per node on Fugaku is 32 GB, it is recommended to limit memory usage to a maximum of around 25 GB, accounting for the memory reserved by the system. Based on our evaluation of the matrix elements in Fortran arrays and the MPI work areas used in this program, the maximum matrix dimension that can be processed directly on a single node is slightly over 30,000. During the development phase, we considered dynamic memory allocation. However, we found that when memory usage approached its limits, it occasionally led to segmentation faults or unexpected numerical results, posing a risk of providing incorrect data to users. To ensure computational reliability, we opted for a fixed-dimension design, capping total memory usage at under 20 GB as a safety margin. This allocation includes approximately 4 GB each for the overlap, kinetic energy, and potential matrices, as well as the MPI work area, with an additional 3 to 4 GB reserved for intermediate work arrays.

Although the final five-body calculations in this study exceed this 20 GB limit, we addressed this by partitioning the calculations and storing the intermediate results in storage. The final diagonalization was then performed on a server equipped with a large-capacity memory for simplicity. Although an exact time cannot be specified as the post-processing server is shared by multiple users, the calculations in Table 2 typically take several hours. However, for actual calculations involving larger dimensions and repeated diagonalizations, we plan to perform diagonalizations using the parallelized subroutine EigenExa on Fugaku as investigated [25].

3.2 Parallel Speed-up Ratio

In this subsection, we demonstrate the parallel performance of the current code using the speed-up ratio achieved with 120 nodes as an example. Even for the advanced calculations of the ^{4}He pentamer discussed in this paper, all necessary computations for the research can be completed within one to several days (a single calculation takes from several to several tens of minutes), provided that high parallel efficiency is achieved (using 120 nodes as in this subsection).

As previously mentioned, the calculations of these matrix elements are entirely independent of one another, and the computation time for all elements within the same channel is practically identical. This represents an ideal case for parallelization, where the execution time is expected to decrease in inverse proportion to the number of nodes. For instance, in the 120-node calculation shown below, if the negligible delays caused by the MPI communication overhead are ignored, the execution time should ideally be reduced to 1/120 of the time required for a single node (representing a 120-fold speed-up).

In Fig. 3, a speed up ratio of MPI parallelization is displayed. The solid line is the computed speed-up ratio and the dashed line is the ideal line. The number of computed matrix elements is 720600 with channel 1. The value is chosen appropriately so that the computation would finish in approximately 10 min on a single node, thereby avoiding any unnecessary load on the system and causing computational delays for other users. The numerical calculations in this study were performed using the `lang/tcsds-1.2.42` compiler environment on the supercomputer Fugaku. Loading this environment specifically defines the execution paths and versions for the development tools, including Fujitsu Compiler 4.12.2 (Fortran/C/C++/Tool) and its associated mathematical libraries.

The source code was compiled with the Fujitsu Fortran compiler using the following command:

```
mpifrtpx -Kfast -SSL2 -mcmodel=large
```

In this command, `-Kfast` specifies maximum optimization to leverage the performance of the A64FX processor, `-SSL2` enables the use of the mathematical library for efficient numerical calculations, and `-mcmodel=large` sets the memory model to accommodate the large-scale memory requirements of few-body systems.

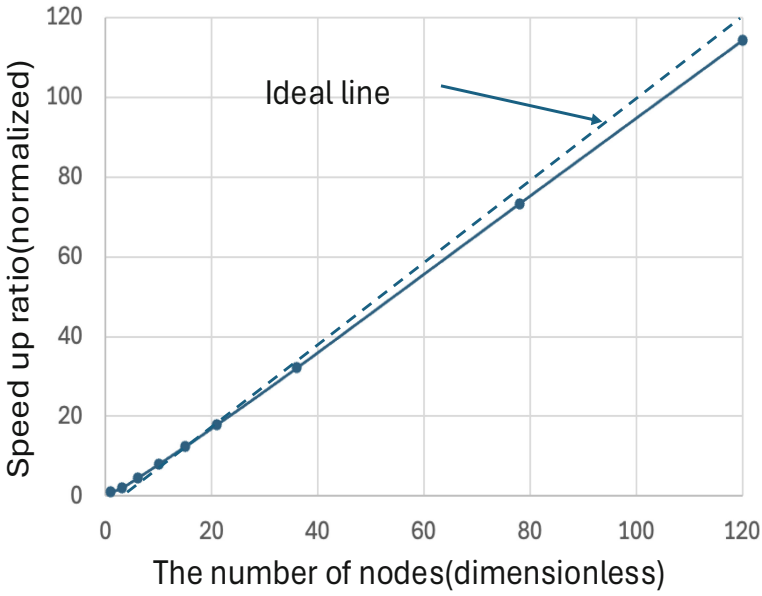


Fig. 3. Parallel speed-up ratio of the matrix-element calculation with MPI parallelization. The code achieves a 118-fold speed-up on 120 nodes (98.3% efficiency).

It is readily apparent that the calculated values are extremely close to the ideal line. The actual measurement yielded a 118-fold speed-up using 120 nodes, which corresponds to an exceptionally high parallel efficiency of approximately 98%. Given that an MPI operation is performed at the final stage to gather all matrix elements into the rank 0, the slight deviation from the ideal 120-fold speed-up is to be expected, considering the inter-node communication involved.

Figure 4 illustrates the speed-up ratio achieved through hybrid parallelization. The solid line represents the computed speed-up ratio, while the dashed line indicates the ideal linear scaling. In this benchmark, a total of 6,481,800 matrix-elements were calculated for Channel 1 for the same reasons as in the MPI implementation. The source code was compiled using the following command:

```
mpifrtpx -Kfast -Kparallel -SSL2 -mcmodel=large
```

In this command, `-Kparallel` is the automatic parallelization option.

The speed-up on 120 nodes reached 122.5-fold, slightly exceeding the ideal linear scaling. While this may appear counterintuitive at first glance, it indicates a so-called "super-linear speed-up" [28]. This phenomenon is likely attributable to improved cache efficiency and optimized memory bandwidth utilization during thread parallelization, as the problem size per node decreased. Specifically, at 120 nodes, the number of matrix-elements to be processed per node is reduced to 1/120 of the original workload, leading to a reduction in the memory occupancy

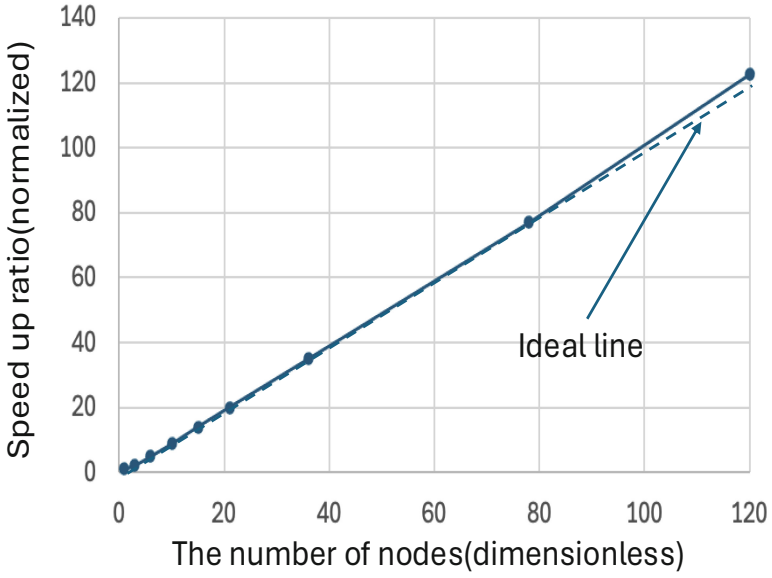


Fig. 4. Parallel speed-up ratio of the matrix-element calculation with Hybrid parallelization. The code achieves a 122.5-fold speed-up on 120 nodes (102% efficiency).

per node. Considering the cache specifications of the A64FX processor (8 MB L2 cache) [29], it is highly probable that the observed super-linear speed-up was driven by an improved cache hit rate. This resulted from partitioning the problem into smaller datasets that fit more efficiently within the cache hierarchy, as expected.

3.3 Computed Energies

In Table 2, the calculated energies and Gaussian range parameters for each channel are shown. The first column shows the channel numbers defined in Table 1. The second column lists the ground-state energies obtained from the calculation in units of Kelvin. The third column shows the energies of the first excited state. While the first excited state is of significant interest in relation to Efimov physics and Bose—Einstein condensation, this paper, as a computational science study, only presents the numerical results for three important channels; physical analysis along with the converged numerical results including other channels will be discussed elsewhere. The subsequent columns list the number of basis functions for each Jacobi coordinate, along with the minimum and maximum values of the Gaussian range parameters. The final column shows the dimension of the matrix that was actually diagonalized. It should be noted that the energy values are not the result of diagonalizing a single channel alone, but the values obtained when the basis functions of the new channel are added to all the basis functions from the preceding rows.

Table 2. Calculated energies and Gaussian range parameters for each channel.

No.	E_0	E_1	N_1	b_1^{min}	b_1^{max}	N_2	b_2^{min}	b_2^{max}	N_3	b_3^{min}	b_3^{max}	N_4	b_4^{min}	b_4^{max}	N	N_{tot}
1	-0.808	-0.440	12	0.5	20	12	0.5	20	12	0.5	20	18	0.7	200	31104	31104
6	-1.205	-0.555	10	0.5	20	10	0.5	20	10	0.5	20	12	0.5	100	12000	43104
11	-1.303	-0.556	8	0.5	20	8	0.5	20	8	0.5	20	10	0.5	40	5120	48224

The large number of basis functions in the first channel is due to its role as the primary channel for reproducing the tetramer system. Furthermore, as previously explained, since a large portion of the other channels is already covered due to the linear dependence of the Gaussian basis, a common practice in GEM is to selectively allocate basis functions to the early channels that are considered physically important. The second channel, Channel 6, represents ^{4}He atoms coupled in an H-type configuration. In the ^{4}He tetramer, this is known as one of the major channels that accelerates convergence when added to the K-type; thus, it is expected to be important for the five-body system as well. The following 3+2 channel physically represents a configuration of a trimer and a dimer, which is a channel that appears for the first time in a five-body system. Therefore, this channel is expected to be important for making the five-body system more deeply bound than the four-body system. Regarding the ground state, other five-body calculations using different interactions have been reported, notably $E = -1.295(14)$ K [30] and $E = -1.300(2)$ K [31]. Since these potentials are tuned to the two-body system, the energies similar to the present calculation are expected for the pentamer. Our result is $E = -1.303$ K, which is consistent with these previous studies.

4 Summary and Conclusions

In this study, we extended a computational code GEM based on the Gauss-Lobe technique—a well-established approach for high-precision few-body calculations—to enable its efficient utilization on modern supercomputers. In particular, our development focused on optimization specifically tailored for hybrid parallel computing environments of the supercomputer Fugaku. Performance benchmarks using the ^{4}He pentamer demonstrated that the modernized GEM code achieves excellent scalability. On 120 nodes, we observed a 122.5-fold speed-up, representing a "super-linear speed-up" that exceeds ideal linear scaling. This result underscores the effectiveness of our hybrid parallelization strategy.

Furthermore, we introduced a practical technique for adjusting the balance between MPI and thread parallelism without necessitating extensive modifications to the original source code. This approach not only ensures the portability and maintainability of the GEM code but also paves the way for efficient precision calculations of the six-body systems or larger. To facilitate a deeper understanding of quantum few-body systems, we plan to release the GEM code to the

public following the preparation of comprehensive documentation and manuals. We anticipate that this modernized GEM framework will serve as a powerful tool for exploring the intricate dynamics of complex few-body systems.

The ^{4}He few-body system is a fundamental quantum-mechanical many-body system that has attracted significant attention, particularly for its relevance to Efimov states and Bose—Einstein condensation. The modernized GEM code developed in this work was implemented using the ^{4}He pentamer as a benchmark, which is at the forefront of few-body physics and for which detailed analysis and understanding are highly anticipated. Consequently, this code is far more than a mere benchmark tool; it is a fully production-ready framework capable of meeting the rigorous demands of frontier research. A comprehensive discussion of the physical properties and detailed analytical results will be reported in a separate, dedicated publication in the near future. Furthermore, we will apply and analyze other five-body systems, such as ^{4}He$+n+n+n+n$ or ^{4}He$+p+n+n+n$ in nulei, which will lead to the future expansion of our methodology to the six-body systems or larger.

Acknowledgements. This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI with Grant Number 25H01270 and by the ERATO TOMOE project with Grant Number JPMJER 2304. We use Supercomputer Fugaku at RIKEN Center for Computational Science (HPCI project ID: hp250328). A part of the development of the GEM code is carried out on a supercomputer Miyabi, provided by the Multidisciplinary Cooperative Research Program in the Center for Computational Sciences, University of Tsukuba.

References

1. Efimov, V.: Phys. Lett. B **33**, 563 (1970)
2. Einstein, A.: Sitzungsber. Preuss. Akad. Wiss, 261 (1924)
3. Anderson, M.H., Ensher, J.R., Matthews, M.R., Wieman, C.E., Cornell, E.A.: Science **269**, 198 (1995)
4. Funaki, Y., Tohsaki, A., Horiuchi, H., Schuck, P., Röpke, G.: Phys. Rev. Lett. **91**, 262502 (2003)
5. Kamimura, M.: Phys. Rev. A **38**, 621 (1988)
6. Kameyama, H., Kamimura, M., Fukushima, Y.: Phys. Rev. C **40**, 974 (1989)
7. Hiyama, E., Kino, Y., Kamimura, M.: Prog. Part. Nucl. Phys. **51**, 223 (2003)
8. O. A. Yakubovsky, Sov. J. Nucl. Phys. **5**, 937 (1967) [Yad. Fiz. **5**, pp.1312 (1967)]
9. Kamada, H., Nogga, A., Glöckle, W., Vary, J.P., Barrett, B.R.: Phys. Rev. C **64**, 044001 (2001)
10. Kievsky, A., Rosati, S., Viviani, M., Marcucci, L.E., Girlanda, L.: J. Phys. G: Nucl. Part. Phys. **35**, 063101 (2008)
11. Navrátil, P., Vary, J.P., Barrett, B.R.: Phys. Rev. Lett. **84**, 5728 (2000)
12. Barrett, B.R., Navrátil, P., Vary, J.P.: Prog. Part. Nucl. Phys. **69**, 131 (2013)
13. Carlson, J., et al.: Rev. Mod. Phys. **87**, 1067 (2015)
14. Hiyama, E.: Few-Body Syst. **53**, 189 (2012)
15. Hiyama, E.: Prog. Theor. Exp. Phys. 01A204 (2012)
16. Varga, K., Suzuki, Y., Lovas, R.G.: Nucl. Phys. A **571**, 447 (1994)

17. Mitroy, J.: Rev. Mod. Phys. **85**, 693 (2013)
18. Aoyama, S., Arai, K., Suzuki, Y., Descouvemont, P., Baye, D.: Few-Body Syst. **52**, 97 (2012)
19. Aoyama, S., Myo, T., Katō, K., Ikeda, K.: Prog. Theor. Phys. **116**, 1 (2006)
20. Aoyama, S., Katō, K., Ikeda, K.: Prog. Theor. Phys. Suppl. **142**, 35 (2001)
21. Suzuki, Y., Varga, K., Usukura, J.: Nucl. Phys. A **631**, 91 (1998)
22. Suzuki, Y., Varga, K.: Stochastic variational approach to quantum-mechanical few-body problems (Lecture notes in physics, Vol. 54), Springer, Berlin (1998)
23. Kukulin, V.I., Krasnopol'sky, V.M.: J. Phys. G: Nucl. Phys. **3**, 79 (1977)
24. Aoyama, S., Myo, T., Yoshida, D.: Computational Science and Its Applications – CSHPC2025 (ICCSA 2025 Workshops), Proceedings, Lecture Notes in Computer Science, Vol. 15890, Springer, pp. 42(2026)
25. Yoshida, D., Hayama, E., Imamura, T., Kanamori, I., Matsufuru, H.: SCA/HPCAsiaWS '26 : Proceedings of the Supercomputing Asia and International Conference on High Performance Computing in Asia Pacific Region Workshops, pp. 133 (2026)
26. Hiyama, E., Kamimura, M.: Front. Phys. **13**(6), 132106 (2018)
27. Naidon, P.: (RIKEN Nishina Centre). private communication
28. Ristov, S., Prodan, R., Gusev, M., Skala, K.: Proceedings of the Federated Conference on Computer Science and Information Systems, p. 889 (2016)
29. Sato, M.: Proceedings of the SC20, No.47, p. 1 (2020)
30. Blume, D., Greene, C.H.: J. Chem. Phys. **112**, 8053 (2000)
31. Bazak, B., Valiente, M., Barnea, N.: Phys. Rev. A **101**, 010501(R) (2020)



Acceleration of Multicomponent Multiple-Precision Arithmetic with Branch-Free Algorithms and SIMD Vectorization

Tomonori Kouya(✉)

Otemon Gakuin University, 2-1-15 Nishiai, Ibaraki, Osaka 567-8502, Japan
t-koya@haruka.otemon.ac.jp
https://www.otemon.ac.jp

Abstract. Multiple-precision floating-point branch-free algorithms can significantly accelerate multicomponent arithmetic implemented by combining hardware-based binary64 and binary32, particularly for triple- and quadruple-precision computations. In this study, we achieved benchmark results on x86 and ARM CPU platforms to quantify the accelerations achieved in linear computations and polynomial evaluations by integrating these algorithms.

Keywords: branch-free · multiple-precision floating-point arithmetic · linear computation · polynomial evaluation

1 Introduction

Recent advances in artificial intelligence (AI), underpinned by deep-learning technologies, have led to the widespread adoption of large language models and generative AI systems, permeating diverse industrial sectors and everyday human activities, including software development and coding. This trend has accelerated the shortening of the mantissa lengths in floating-point formats employed in AI computations to achieve higher performance. However, in scientific and engineering computing, double-precision arithmetic is the minimum standard, and ill-conditioned problems frequently require higher precision. In such cases, the effective use of single-instruction multiple-data (SIMD) instructions and OpenMP-based parallelization techniques is crucial for harnessing the full capabilities of modern hardware. Among the existing multiple-precision arithmetic frameworks, arbitrary-precision arithmetic is dominated by MPFR [15], which relies on the multiple-precision natural-number arithmetic (MPN) kernel of GMP [5]. MPC [3], an arbitrary precision complex arithmetic library built on MPFR that achieves acceleration through the Karatsuba 3M multiplication algorithm, is widely used.

In contrast, fixed-precision multicomponent arithmetic based on error-free transformation (EFT) techniques, which employs combinations of multiple IEEE

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 508–524, 2027.
https://doi.org/10.1007/978-3-032-30521-3_32

binary32 (24-bit mantissa and 32-bit total length) and binary64 values supported by existing hardware, is represented by quadruple-double (QD) [1] for CPUs and GQD [12]—a CUDA port of QD, though dated—for GPUs. When a relatively short mantissa length suffices, such as in double-double (DD, 106-bit), triple-double (TD, 159-bit), and quadruple-double (QD, 212-bit) arithmetic implemented by combining two to four binary64 values, computations can be executed faster than with MPFR. Kotakemori et al.'s Lis [7] and Hishinuma et al. [6] further demonstrated that basic linear computations can be accelerated by leveraging AVX, AVX2, and AVX-512, which are x86-based SIMD instruction sets. Building on an existing AVX2 implementation, we demonstrated that basic linear computations in DD, TD, and QD precision can be accelerated on the Raspberry Pi and Snapdragon platforms by exploiting the Neon instructions [18].

Nevertheless, DD, TD, and QD arithmetic require renormalization at the conclusion of each operation. In particular, for TD and QD arithmetic, this step has been identified as a bottleneck that impedes performance gains through SIMD vectorization. Double-precision pair arithmetic, which bypasses renormalization, has been proposed as a faithfully rounded pair-arithmetic scheme [11], and practical sparse matrix solver implementations that exploit this efficiency have been reported [14]. The branch-free algorithms for multiple-precision floating-point arithmetic addressed in this work [17] eliminate all conditional branches from every algorithmic step, including renormalization, and reduce the operation count for TD and QD arithmetic relative to conventional algorithms. Consequently, this methodology can substantially accelerate triple- and quadruple-precision multicomponent arithmetic implemented by combining hardware-based binary64 and binary32 values and is immediately applicable to existing fixed-precision multiple-precision arithmetic implementations. Because the benchmarks presented in the original paper provided an insufficient comparison against existing SIMD implementations, we first quantified the speedups achievable in basic linear computations by incorporating these algorithms on the x86 and ARM CPU platforms. Furthermore, as a practical application, we analyzed the effectiveness of the proposed approach by applying it to a multiple-precision algebraic equation solver.

2 Branch-Free Algorithms for DD, TD, and QD Addition and Multiplication

The extension of fixed-precision floating-point numbers via EFT was pioneered by Dekker [2]. Multiple-precision arithmetic with an extended mantissa length was implemented by employing EFT primitives, represented by QuickTwoSum (Algorithms 1, 2, 3), which return both a rounded result and its rounding error using only existing floating-point operations (addition $\oplus$, subtraction $\ominus$, and multiplication $\otimes$). In the following, all variables are assumed to be exactly representative as normalized floating-point numbers.

QuickTwoSum (Algorithm 1) returns the floating-point addition result $s = a \oplus b$ and its rounding error e, but only under the precondition $|a| \geq |b|$; if no overflow occurs, it maintains the relation $a + b = s + e$.

Algorithm 1 $(s, e) :=$ QuickTwoSum(a, b)

$s := a \oplus b$; $e := b \ominus (s \ominus a)$; **return** (s, e)

TwoSum (Algorithm 2) returns the floating-point addition result $s = a \oplus b$ and its rounding error e unconditionally, regardless of the relative magnitudes of a and b, while maintaining the relationship $a + b = s + e$.

Algorithm 2 $(s, e) :=$ TwoSum(a, b)

$s := a \oplus b$; $v := s \ominus a$; $e := (a \ominus (s \ominus v)) \oplus (b \ominus v)$; **return**$(s, e)$

TwoProd (Algorithm 3) returns the floating-point multiplication result $p = a \otimes b$ and its rounding error e while maintaining the relationship $ab = p + e$. Because virtually all floating-point hardware currently in operation supports fused multiply–add (FMA) instructions, the FMA-based variant of TwoProd was employed throughout this work.

Algorithm 3 $(p, e) :=$ TwoProd(a, b)

$p := a \otimes b$; $e := \mathrm{FMA}(a, b, -p)$ $(= a \times b - p)$; **return**(p, e)

As these EFT operations are constructed entirely from existing floating-point operations and contain no conditional branches, they can be implemented directly using SIMD instructions. With AVX2, which employs 256-bit registers on x86 architectures, four binary64 or eight binary32 values can be processed simultaneously by EFT operations. In contrast, with Neon on AArch64, which employs 128-bit registers, two binary64 or four binary32 values can be processed simultaneously. Computations that contain no conditional branches and can be executed without overflowing SIMD registers in this manner are termed branch-free (BF).

By combining EFT operations, multiple-precision arithmetic with extended mantissa can be implemented by composing multiple floating-point numbers; this is referred to as the multicomponent (or multiterm) method. Because subtraction reduces to sign negation and division can be implemented via Newton's method that combines addition and multiplication, these two operations are crucial from a performance perspective. Correct handling of the rounding errors is critical in this implementation and requires a carefully derived rigorous error analysis. As

the mantissa length increases, the operation count increases, and error management becomes increasingly complex. Zhang & Aiken proposed an abstract model for automated reasoning that simplifies floating-point arithmetic [16] and introduced branch-free multicomponent multiple-precision addition and multiplication algorithms. In the following section, we present and discuss BF algorithms.

2.1 Remarks on Double-Word (DW) Arithmetic

The simplest form of multicomponent multiple-precision floating-point arithmetic is the DW arithmetic. When binary64 (double-precision) is used as the base format, it is abbreviated as DD; when binary32 (single-precision) is used, it is abbreviated as DS. Algorithm 4 presents the default DW addition defined in QD [1], which is the sloppy or fast variant; its final rounding error is larger than that of the accurate version (Algorithm 5).

Algorithm 4 $c[2] := \mathrm{DWAdd}(a[2],\ b[2])$

$(s, e) := \mathrm{TwoSum}(a[0], b[0])$; $e := e \oplus (a[1] \oplus b[1])$; $(c[0], c[1]) := \mathrm{QuickTwoSum}(s, e)$;
return$(c[2])$

Algorithm 5 $c[2] := \mathrm{DWAccurateAdd}(a[2],\ b[2])$

$(s_1, s_2) := \mathrm{TwoSum}(a[0], b[0])$; $(t_1, t_2) := \mathrm{TwoSum}(a[1], b[1])$;
$s_2 := s_2 \oplus t_1$; $(s_1, s_2) := \mathrm{QuickTwoSum}(s1, s2)$;
$s_2 := s_2 \oplus t_2$; $(c[0], c[1]) := \mathrm{QuickTwoSum}(s1, s2)$;
return$(c[2])$

Zhang and Aiken proposed Algorithm 6 as a branch-free reformulation for accurate DW addition.

Algorithm 6 $c[2] := \mathrm{DWBFAdd}(a[2],\ b[2])$

$(g_1, g_{1e}) := \mathrm{TwoSum}(a[0], b[0])$;
$(g_2, g_{2e}) := \mathrm{TwoSum}(a[1], b[1])$;
$(g_3, g_{3e}) := \mathrm{QuickTwoSum}(g_1, g_2)$;
$g_4 := g_{1e} \oplus g_{2e}$;
$g_5 := g_4 \oplus g_{3e}$;
$(c[0], c[1]) := \mathrm{QuickTwoSum}(g_3, g_5)$;
return$(c[2])$

Algorithm 7 presents the DW multiplication defined as the default in QD.

Algorithm 7 $c[2] := \mathrm{DWMul}(a[2], b[2])$

$(p_1, p_2) := \mathrm{TwoProd}(a[0], b[0])$; $p_2 := p_2 \oplus (a[0] \otimes b[1] \oplus a[1] \otimes b[0])$;
$(c[0], c[1]) := \mathrm{QuickTwoProd}(p_1, p_2)$;
return$(c[2])$

Algorithm 8 $c[2] := \mathrm{DWBFMul}(a[2], b[2])$

$(p_{00}, p_{e00}) := \mathrm{TwoProd}(a[0], b[0])$;
$p_{01} := a[0] \otimes b[1]$;
$p_{10} = a[1] \otimes b[0]$;
$g_1 := p_{01} \oplus p_{10}$;
$g_2 := p_{e00} \oplus g_1$;
$(c[0], c[1]) := \mathrm{QuickTwoSum}(p_{00}, g_2)$;
return$(c[2])$

The corresponding BF multiplication is presented as Algorithm 8. This algorithm can be interpreted as performing all necessary multiplications, followed by arranging the error-handling steps in a branch-free manner.

Consequently, the BF variants of DW addition and multiplication involve no fewer floating-point operations than their conventional counterparts. In addition, as discussed in the subsequent sections, they exhibit no measurable performance benefit under the current default DW arithmetic. Therefore, in the following benchmark tests, DD arithmetic employs conventional (non-BF) implementations unless otherwise stated; only TD and QD arithmetic utilize the BF algorithms presented hereafter.

2.2 Remarks on TW Arithmetic

For triple-precision arithmetic (triple-word, TW), one approach is to consider the fourth component of the QD arithmetic as zero. In contrast, Fabiano et al. [4] proposed a formulation with a reduced operation count. Through SIMD implementation, we validated that the former approach outperforms the latter [9]; therefore, Algorithm 9 is used for TW addition and Algorithm 10 for TW multiplication.

Both algorithms conclude with a call to the TW renormalization function TWRenormalize [4] [1], which requires up to five QuickTwoSum operations and necessitates conditional branching via if-statements. This function has not yet been SIMD-vectorized and is considered the primary source of performance degradation.

By contrast, the proposed BF algorithms presented in Algorithms 11 and 12 introduce no conditional branches in the renormalization-equivalent steps and require fewer operations than their conventional counterparts. This renders them highly amenable to performance improvements through SIMD vectorization.

As listed in Table 1, the branch-free triple-word multiplication has the least total number of primary computations compared with our currently adopted

Algorithm 9 $c[3] := \mathrm{TWAdd}(a[3], b[3])$

$s_0 := a[0] \oplus b[0]$; $s_1 := a[1] \oplus b[1]$; $s_2 := a[2] \oplus b[2]$;
$v_0 := s_0 \ominus a[0]$;$v_1 := s_1 \ominus a[1]$;$v_2 := s_2 \ominus a[2]$;
$u_0 := s_0 \ominus v_0$; $u_1 := s_1 \ominus v_1$; $u_2 := s_2 \ominus v_2$;
$w_0 := a[0] \ominus u_0$;$w_1 := a[1] \ominus u_1$; $w_2 := a[2] \ominus u_2$;
$u_0 := b[0] \ominus v_0$; $u_1 := b[1] \ominus v_1$; $u_2 := b[2] \ominus v_2$;
$t_0 := w_0 \oplus u_0$; $t_1 := w_1 \oplus u_1$; $t_2 := w_2 \oplus u_2$;
$(s_1, t_0) := \mathrm{TwoSum}(s_1, t_0)$; $(i_1, i_2) := \mathrm{TwoSum}(s_2, t_0)$;
$(s_2, i_3) := \mathrm{TwoSum}(t_1, i_1)$; $(t_0, t_1) := \mathrm{TwoSum}(i_2, i_3)$ $t_0 := t_0 \oplus t_1 \oplus t_2$;
$c[3] := \mathrm{TWRenormalize}(s_0, s_1, s_2, t_0)$;
return$(c[3])$

Algorithm 10 $c[3] := \mathrm{TWMul}(a[3], b[3])$

$(p_0, q_0) := \mathrm{TwoProd}(a[0], b[0])$; $(p_1, q_1) := \mathrm{TwoProd}(a[0], b[1])$;
$(p_2, q_2) := \mathrm{TwoProd}(a[1], b[0])$;
$(p_3, q_3) := \mathrm{TwoProd}(a[0], b[2])$; $(p_4, q_4) := \mathrm{TwoProd}(a[1], b[1])$;
$(p_5, q_5) := \mathrm{TwoProd}(a[2], b[0])$;
$(i_1, i_2) := \mathrm{TwoSum}(p_1, p_2)$; $(p_1, i_3) := \mathrm{TwoSum}(q_0, i_1)$; $(p_2, q_0) := \mathrm{TwoSum}(i_2, i_3)$;
$(i_1, i_2) := \mathrm{TwoSum}(p_2, q_1)$; $(p_2, i_3) := \mathrm{TwoSum}(q_2, i_1)$; $(q_1, q_2) := \mathrm{TwoSum}(i_2, i_3)$;
$(i_1, i_2) := \mathrm{TwoSum}(p_3, p_4)$; $(p_3, i_3) := \mathrm{TwoSum}(p_5, i_1)$; $(p_4, p_5) := \mathrm{TwoSum}(i_2, i_3)$;
$(s_0, t_0) := \mathrm{TwoSum}(p_2, p_3)$; $(s_1, t_1) := \mathrm{TwoSum}(q_1, p_4)$; $s_2 := q_2 \oplus p_5$;
$(s_1, t_0) := \mathrm{TwoSum}(s_1, t_0)$; $s_2 := s_2 \oplus (t_0 \oplus t_1)$; $(q_0, q_3) := \mathrm{TwoSum}(q_0, q_3)$;
$(q_4, q_5) := \mathrm{TwoSum}(q_4, q_5)$; $(t_0, t_1) := \mathrm{TwoSum}(q_0, q_4)$; $t_1 := t_1 \oplus (q_3 \oplus q_5)$;
$(t_0, t_1) := \mathrm{TwoSum}(q_3, s_1)$;
$c[3] := \mathrm{TWRenormalize}(p_0, p_1, s_0, t_0)$;
return$(c[3])$

and Fabiano's and triple-word multiplications. However, we could not find a clear difference among the three types of triple-word additions.

2.3 Remarks on QW Arithmetic

Quadruple-precision arithmetic (quadruple-word, QW) is based on the implementation provided in QD. As noted previously, TD arithmetic is an abbreviated form of QD arithmetic with the fourth component zeroed. Both the QD addition and multiplication conclude with a call to the QW renormalization function (QWRenormalize [1]), which contains conditional branches and requires up to seven QuickTwoSum function calls.

The BF algorithms are presented in Algorithms 13 and 14. Although the time required to discover these BF algorithms is unknown, the resulting algorithms are remarkably concise and involve substantially reduced operation counts. Consequently, the benefit of SIMD vectorization is expected to be more pronounced for QW arithmetic than for TW arithmetic.

We present Table 2 for comparison with our adopted and branch-free quadruple-word algorithms. Similarly, as listed in Table 1, branch-free multi-

Algorithm 11 $c[3] := \mathrm{TWBFAdd}(a[3], b[3])$

$(a_1, b_1) := \mathrm{TwoSum}(a[0], b[0])$; $(c_1, d_1) := \mathrm{TwoSum}(a[1], b[1])$;
$(e_1, f_1) := \mathrm{TwoSum}(a[2], b[2])$;
$(a_2, c_2) := \mathrm{QuickTwoSum}(a_1, c_1)$; $b_2 := b_1 \oplus f_1$;
$(d_2, e_2) := \mathrm{TwoSum}(d_1, e_1)$; $(a_3, d_3) := \mathrm{QuickTwoSum}(a_2, d_2)$;
$(b_3, c_3) := \mathrm{TwoSum}(b_2, c_2)$; $c_4 := c_3 \oplus e_2$;
$(c_5, d_5) := \mathrm{TwoSum}(c_4, d_3)$; $(b_6, c_6) := \mathrm{TwoSum}(b_3, c_5)$;
$(c[0], b_7) = \mathrm{QuickTwoSum}(a_3, b_6)$; $c_7 := c_6 \oplus d_5$;
$(c[1], c[2]) := \mathrm{QuickTwoSum}(b_7, c_7)$;
return$(c[3])$

Algorithm 12 $c[3] := \mathrm{TWBFMul}(a[3], b[3])$

$(a_0, b_0) := \mathrm{TwoProd}(a[0], b[0])$; $(c_0, e_0) := \mathrm{TwoProd}(a[0], b[1])$;
$(d_0, f_0) := \mathrm{TwoProd}(a[1], b[0])$;
$g_0 := a[0] \otimes b[2]$; $h_0 := a[1] \otimes b[1]$; $i_0 := a[2] \otimes b[0]$;
$(c_1, d_1) := \mathrm{TwoSum}(c_0, d_0)$; $e_1 := e_0 \oplus f_0$; $g_1 := g_0 \oplus i_0$;
$(b_2, c_2) := \mathrm{TwoSum}(b_0, c_1)$; $g_2 := g_1 \oplus h_0$; $(a_3, b_3) := \mathrm{QuickTwoSum}(a_0, b_2)$;
$c_3 := c_2 \oplus d_1$; $e_3 := e_1 \oplus g_2$; $c_4 := c_3 \oplus e_3$;
$(b_5, c_5) := \mathrm{QuickTwoSum}(b_3, c_4)$; $(c[0], b_6) := \mathrm{QuickTwoSum}(a_3, b_5)$;
$(c[1], c[2]) := \mathrm{QuickTwoSum}(b_6, c_5)$;
return$(c[3])$

plication is the most lightweight, and no significant difference exists between the two types of quadruple-word additions.

3 Benchmark Tests for Real and Complex Square Matrix Multiplication

As discussed previously, the adoption of BF arithmetic is expected to yield significant performance gains through SIMD vectorization for triple- and quadruple-precision computations. Here, we report performance evaluation results for real and complex square matrix multiplication obtained by integrating BF arithmetic

Table 1. Operation counts for Triple-Word arithmetic

Triple-Word (TW)	Add	Sub	Mul	Div	FMA
TW+TW	19	34	0	0	0
Fabiano TW+TW	20	40	0	0	0
BF TW+TW	21	36	0	0	0
TW×TW	42	74	6	0	6
Fast Fabiano TW×TW	24	38	0	0	3
BF TW×TW	14	16	6	0	3

Algorithm 13 $c[4] := \mathrm{QWBFAdd}(a[4], b[4])$

$(a_1, b_1) := \mathrm{TwoSum}(a[0], b[0])$; $(c_1, d_1) := \mathrm{TwoSum}(a[1], b[1])$;
$(e_1, f_1) := \mathrm{TwoSum}(a[2], b[2])$;
$(g_1, h_1) := \mathrm{TwoSum}(a[3], b[3])$; $(a_2, c_2) := \mathrm{QuickTwoSum}(a_1, c_1)$;
$b_2 := b_1 \oplus h_1$; $(d_2, e_2) := \mathrm{TwoSum}(d_1, e_1)$;
$(f_2, g_2) := \mathrm{TwoSum}(f_1, g_1)$; $(b_3, g_3) := \mathrm{TwoSum}(b_2, g_2)$;
$(c_3, d_3) := \mathrm{QuickTwoSum}(c_2, d_2)$; $(e_3, f_3) := \mathrm{TwoSum}(e_2, f_2)$;
$(a_4, c_4) := \mathrm{QuickTwoSum}(a_2, c_3)$; $(d_4, e_4) := \mathrm{QuickTwoSum}(d_3, e_3)$;
$(b_5, d_5) := \mathrm{TwoSum}(b_3, d_4)$; $e_5 := e_4 \oplus f_3$;
$(b_6, c_6) := \mathrm{TwoSum}(b_5, c_4)$; $(d_6, e_6) := \mathrm{TwoSum}(d_5, e_5)$;
$(a_7, b_7) := \mathrm{QuickTwoSum}(a_4, b_6)$; $(c_7, d_7) := \mathrm{QuickTwoSum}(c_6, d_6)$;
$e_8 := e_6 \oplus g_3$; $(b_8, c_8) := \mathrm{QuickTwoSum}(b_7, c_7)$;
$d_9 := d_7 \oplus e_8$; $(c[0], b_{10}) := \mathrm{QuickTwoSum}(a_7, b_8)$;
$(c_{10}, d_{10}) := \mathrm{QuickTwoSum}(c_8, d_9)$; $(c[1], c_{11}) := \mathrm{QuickTwoSum}(b_{10}, c_{10})$;
$(c[2], c[3]) := \mathrm{QuickTwoSum}(c_{11}, d_{10})$;
return$(c[4])$

Algorithm 14 $c[4] := \mathrm{QDBFMul}(a[4], b[4])$

$(a_0, b_0) := \mathrm{TwoProd}(a[0], b[0])$; $(c_0, e_0) := \mathrm{TwoProd}(a[0], b[1])$;
$(d_0, f_0) := \mathrm{TwoProd}(a[1], b[0])$; $(g_0, j_0) := \mathrm{TwoProd}(a[0], b[2])$;
$(h_0, k_0) := \mathrm{TwoProd}(a[1], b[1])$ $(i_0, l_0) := \mathrm{TwoProd}(a[2], b[0])$;
$m_0 := a[0] \otimes b[3]$; $n_0 := a[1] \otimes b[2]$; $o_0 := a[2] \otimes b[1]$;
$p_0 := a[3] \otimes b[0]$; $(c_1, d_1) := \mathrm{TwoSum}(c_0, d_0)$;
$(e_1, f_1) := \mathrm{TwoSum}(e_0, f_0)$; $(g_1, i_1) := \mathrm{TwoSum}(g_0, i_0)$;
$j_1 := j_0 \oplus l_0$; $m_1 := m_0 \oplus p_0$; $n_1 := n_0 \oplus o_0$;
$(b_2, c_2) := \mathrm{TwoSum}(b_0, c_1)$; $(e_2, h_2) := \mathrm{TwoSum}(e_1, h_0)$;
$f_2 := f_1 \oplus j_1$; $i_2 := i_1 \oplus k_0$; $m_2 := m_1 \oplus n_1$;
$(a_3, b_3) := \mathrm{QuickTwoSum}(a_0, b_2)$; $(c_3, d_3) := \mathrm{QuickTwoSum}(c_2, d_1)$;
$(e_3, g_3) := \mathrm{TwoSum}(e_2, g_1)$; $f_3 := f_2 \oplus m_2$; $h_3 := h_2 \oplus i_2$;
$(c_4, e_4) := \mathrm{TwoSum}(c_3, e_3)$; $d_4 := d_3 \oplus h_3$; $f_4 := f_3 \oplus g_3$;
$d_5 := d_4 \oplus e_4$; $(c_6, d_6) := \mathrm{TwoSum}(c_4, d_5)$; $(b_7, c_7) := \mathrm{TwoSum}(b_3, c_6)$;
$d_7 := d_6 \oplus f_4$; $(c[0], b_8) := \mathrm{QuickTwoSum}(a_3, b_7)$;
$(c_8, d_8) := \mathrm{TwoSum}(c_7, d_7)$; $(c[1], c_9) := \mathrm{TwoSum}(b_8, c_8)$;
$(c[2], c[3]) := \mathrm{QuickTwoSum}(c_9, d_8)$;
return$(c[4])$

into our multiple-precision basic linear computation library BNCmatmul [8], executed on the following x86 AVX2 and Arm Neon environments.

EPYC AMD EPYC 9354P 3.7 GHz 32 cores, Ubuntu 20.04.6 LTS, Intel Compiler version 2021.10.0, GNU MP 6.2.1, MPFR 4.1.0, MPC 1.2.1

Snapdragon Lenovo IdeaPad Slim 5 (Snapdragon X Plus X1P-42-100, 3.4 GHz, 34 MB L3 cache, 8 cores), Windows 11, WSL2 + Ubuntu 24.04.2 LTS, GCC 13.3.0

Standard intrinsic functions, provided as C functions, were used for SIMD implementation in both environments. Arm Neon vectorization and OpenMP

Table 2. Operation counts for quad-word arithmetic

Quad-Word (QW)	Add	Sub	Mul	Div	FMA
QW+QW	26	58	0	0	0
BF QW+QW	37	66	0	0	0
QW×QW	67	94	12	0	9
BF QW×QW	38	52	10	0	6

parallelization were implemented primarily with assistance from the Anthropic Claude by converting the existing AVX2 code [18].

3.1 Real Matrix Multiplication

The square matrices employed in the tests are defined as follows:

$$\begin{aligned} &C := AB, \\ &A = [a_{ij}]_{i,j=1,2,\ldots,n}, B = [b_{ij}]_{i,j=1,2,\ldots,n} \in \mathbb{R}^{n\times n} \\ &a_{ij} := \sqrt{5}(i+j-1), b_{ij} := \sqrt{3}(n-i+1) \end{aligned} \tag{1}$$

as defined in (1).

Our library supports four matrix multiplication schemes: the naive triple-loop method, a block-based algorithm, Strassen's algorithm, and the Ozaki scheme. In this study, Strassen's algorithm was employed to isolate the performance contribution of the BF arithmetic. The recursion in Strassen's algorithm terminates when the submatrix size decreases below 32×32. The performance benefit of BF arithmetic was measured for both nonvectorized and SIMD-vectorized (AVX2 and Neon) implementations. Execution times with OpenMP parallelization using 8 and 32 threads were also reported. Because our Strassen routine did not achieve further speedup beyond eight threads, the results for the block-based algorithm (B32 threads) were included only for the 32-thread case. The numerical accuracy of all computed results falls within the ranges of 29.0–30.4 significant decimal digits for DD, 45.5–47.5 for TD, and 61.7–63.5 for QD precision.

Figure 1 shows the computation times for real square matrix multiplication at TD and QD precision. The horizontal axis represents the matrix size $n = 32$–2049, and the vertical axis represents the computation time in seconds.

Notably, BF arithmetic provides no benefit and degrades performance for DD precision, whereas it yields clear speedups for TD and QD precisions. The resulting speedups over the MPBLAS [13] DD- and QD-precision Rgemm for $n = 2049$ are summarized in Table 3.

Table 4 lists the average speedup ratios owing to BF arithmetic for DD precision, where BF arithmetic offers no benefit compared to nonSIMD (NS) and SIMD-vectorized implementations.

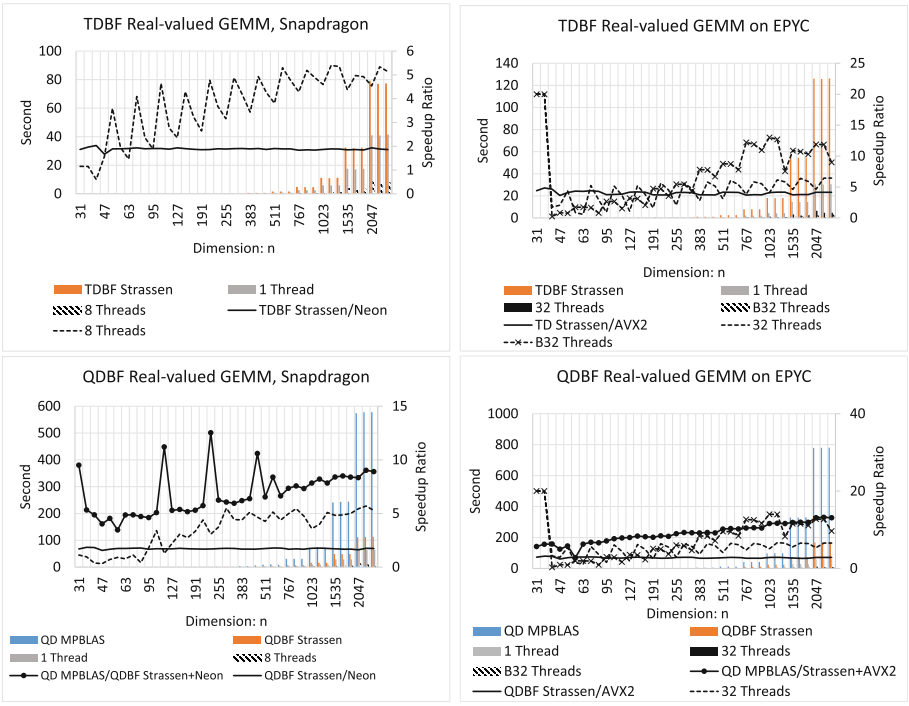


Fig. 1. Computation time (s) of real-valued Strassen matrix multiplication: Snapdragon (left), EPYC (right).

Table 3. Real-valued matrix multiplication time (s) for $n = 2049$

	MPBLAS	S+SIMD	BF S+SIMD	8 Threads	32 Threads	B32 Threads
DD Snapdragon	13.6	9.7	15.4	2.5	–	–
EPYC	13.9	6.7	10.3	1.5	1.4	0.9
TD Snapdragon	–	61.7	39.9	8.1	–	–
EPYC	–	46.1	30.4	4.8	4.7	3.4
QD Snapdragon	577.5	101.4	65.6	12.3	–	–
EPYC	781.0	117.2	59.9	9.2	9.1	6.2

Table 4. Average speedup with branch-free (BF) arithmetic: standard / BF

Speedup	DD		TD		QD	
ratio	NS	SIMD	NS	SIMD	NS	SIMD
Snapdragon	0.6	0.7	1.2	1.4	2.0	1.4
EPYC	1.0	0.7	1.4	1.5	1.5	1.9

As expected, DD precision exhibited a slowdown of 0.6–1.0×; however, TD precision achieved speedups of 1.2–1.5× and QD precision achieved speedups of 1.4–2.0×. Although the true benefit of BF arithmetic is difficult to discern from the benchmarks in the original study, the aforementioned results demonstrate that BF arithmetic effectively accelerates TD and QD precision computations. Furthermore, speedups were achieved through Arm Neon vectorization and OpenMP-based parallelization.

3.2 Benchmark Tests for Complex Matrix Multiplication

Owing to the previously demonstrated superiority of BF arithmetic for TD and QD computations, complex matrix multiplication benchmarks employed BF-based implementations for comparison. Complex square matrix multiplication is defined using a uniform random variable $r_u \in [0,1]$ and a standard normal random variable $r_n \in \mathbb{R}$ as

$$\begin{aligned} C &:= AB, \\ A = [a_{ij}]_{i,j=1,2,\ldots,n}, B &= [b_{ij}]_{i,j=1,2,\ldots,n} \in \mathbb{C}^{n\times n} \\ a_{ij} &:= \exp(r_n)(r_u - 1/2) + \exp(r_n)(r_u - 1/2)\sqrt{-1} \\ b_{ij} &:= \exp(r_n)(r_u - 1/2) + \exp(r_n)(r_u - 1/2)\sqrt{-1} \end{aligned} \tag{2}$$

Our complex matrix multiplication was implemented using three real matrix multiplications via the 3M method, a variant of the Karatsuba algorithm [10] as follows:

$$\begin{aligned} C_1 := \mathrm{Re}(A)\mathrm{Re}(B),\ C_2 &:= \mathrm{Im}(A)\mathrm{Im}(B) \\ \mathrm{Re}(C) &:= C_1 - C_2 \\ \mathrm{Im}(C) &:= (\mathrm{Re}(A) + \mathrm{Im}(A))(\mathrm{Re}(B) + \mathrm{Im}(B)) - C_1 - C_2. \end{aligned} \tag{3}$$

Therefore, both the Strassen- and block-based algorithms perform three real matrix multiplications. The numerical accuracy, measured as the minimum of the real and imaginary parts, falls within the ranges of 23.2–24.4 significant decimal digits for DD, 39.2–40.8 for TD, and 55.8–56.5 for QD precision.

Figure 2 presents the benchmark results for complex matrix multiplication on the EPYC and Snapdragon platforms.

The results demonstrate performance characteristics consistent with those observed for real matrix multiplication. In particular, extracting the computation times for $n = 2049$ reveals that our implementation achieved speedups that significantly exceeded those for real matrix multiplication relative to MPBLAS, as listed in Table 5.

Consequently, our complex matrix multiplication requires only 2.8–3.5× the computation time of real matrix multiplication across all precision levels. In contrast, MPBLAS, which employs the standard 4M method for each operation, requires 3.5–18.4× the computation time. These results indicate that the 3M-based implementation becomes increasingly advantageous as precision and matrix size increase.

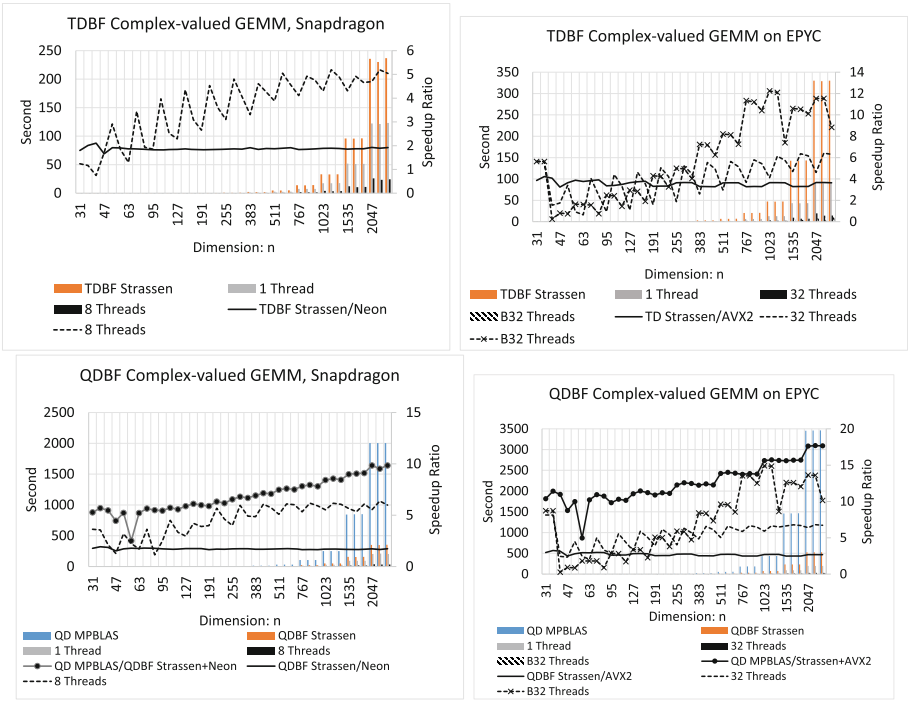


Fig. 2. Computation time (s) of complex-valued Strassen matrix multiplication: Snapdragon (left), EPYC (right).

4 Benchmark Tests for Polynomial Evaluation and Algebraic Equation Solvers

As demonstrated by the matrix multiplication benchmarks, BF arithmetic offers no speedup in DD precision when conventional sloppy addition and multiplication are employed. In this section, we present the benchmark results for a polynomial evaluation and an algebraic equation solver based on the Durand–Kerner (DK) method, both of which utilize TD and QD precision arithmetic, for which performance improvements are anticipated. As described previously, we did not use the BF algorithm for the DD evaluation of a polynomial, but used the BF polynomial evaluation for TD and QD arithmetic.

4.1 Evaluation of Real-Coefficient Polynomial Functions

For the polynomial evaluation, the coefficients of the degree-n real-coefficient polynomial $p_n(x) = \sum_{i=0}^{n} a_i x^i$ were generated randomly, and the results were compared among the standard Horner's method, Estrin's method, and SIMD-vectorized Estrin's method.

Table 5. Complex-valued matrix multiplication time (s) for $n = 2049$

Precision	Platform	MPBLAS	S+SIMD	8 Threads	32 Threads	B32 Threads
DD	Snapdragon	72.1	29.9	7.8	–	–
	EPYC	254.3	20.4	5.0	4.4	3.0
TDBF	Snapdragon	–	123.2	24.4	–	–
	EPYC	–	90.5	14.6	14.4	10.3
QDBF	Snapdragon	2004.2	203.4	34.0	–	–
	EPYC	3463.1	196.1	29.2	29.2	19.4

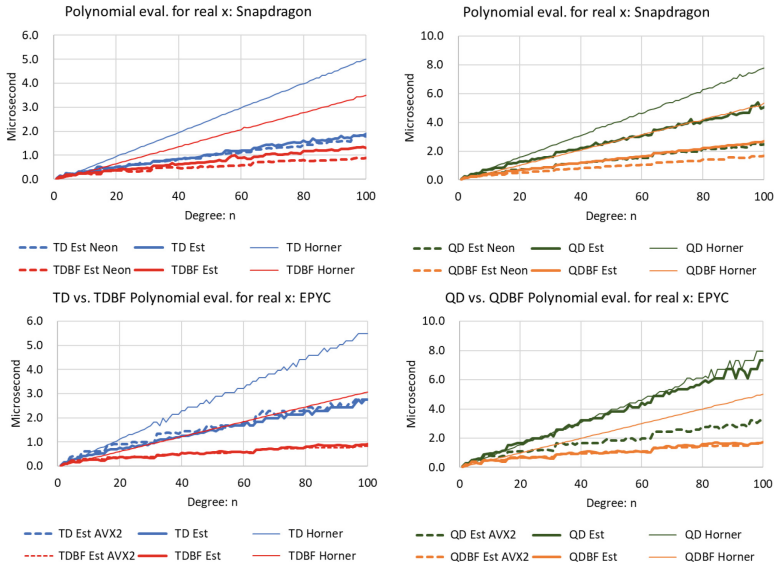


Fig. 3. Computation time (μs) for evaluating a real-coefficient polynomial $p_n(x)$ at real arguments: Snapdragon (top), EPYC (bottom).

Figure 3 shows the computation time (in microseconds) required to evaluate the real-coefficient polynomial $p_n(x)$ in real arguments. The upper panel corresponds to the Snapdragon platform and the lower panel to the EPYC platform; the standard arithmetic (NoBF) is compared with branch-reduction BF arithmetic (BF). As the degree (or equivalently, the evaluation count corresponding to the problem size) increases, the computation time increases, and the effects of branch elimination and vectorization become more pronounced, particularly for high-precision variants such as TD and QD, where the operation count and cost are dominant. Across both architectures, a consistent trend is observed: BF reduces the computation time relative to NoBF for TD and QD precisions.

Figure 4 shows the computation time (in microseconds) required to evaluate the real-coefficient polynomial $p_n(x)$ for the complex arguments. Because each

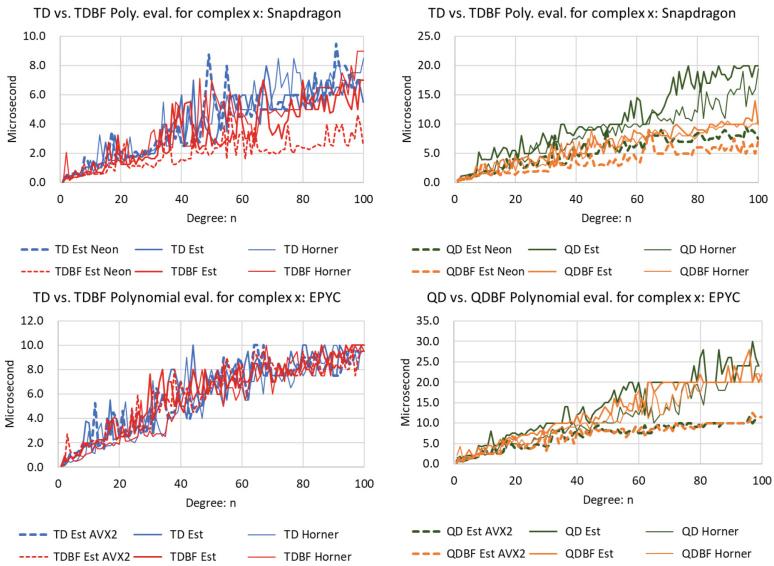


Fig. 4. Computation time (μs) for evaluating a real-coefficient polynomial $p_n(x)$ at complex arguments: Snapdragon (top) and EPYC (bottom).

complex multiplication and addition decomposes into multiple real operations, the computation times are higher than those for real-argument evaluation, as shown in Fig. 3. In contrast, as the number of operations increases, the impact of the conditional branching and normalization steps also increases; therefore, BF arithmetic should, in principle, remain effective for complex-argument evaluation. Although the Snapdragon results largely conform to this expectation, the EPYC results show a slight discernible difference. In particular, the performance difference for TD arithmetic appears to be minimal, even when one exists.

4.2 Algebraic Equation Solver Based on the DK Method

The method being compared is the quadratically convergent DK method, which is a classical simultaneous iterative solver. The approximate roots after k iterations are expressed as follows:

$$\mathbf{z}_k = [z_1^{(k)} \; z_2^{(k)} \; \ldots \; z_n^{(k)}]^T \in \mathbb{C}^n,$$

The recurrence relationship, represented using the monic polynomial $q_n(x) = p_n(x)/a_n$, is expressed by

$$z_i^{(k+1)} := z_i^{(k)} - \frac{q_n(z_i^{(k)})}{\prod_{j=1, j\neq i}^{n}(z_i^{(k)} - z_j^{(k)})} \tag{4}$$

as the quadratically convergent DK method.

The Aberth initial values are expressed as

$$z_i^{(0)} := -\frac{c_{n-1}}{n} + r \exp\left\{\left(\frac{2(i-1)\pi}{n} + \frac{3}{2n}\right)\mathrm{i}\right\} \tag{5}$$

In the experiments, the radius r was defined using the number of nonzero coefficients $n_{\mathrm{nz}} \le n$ as

$$r := \max_{0 \le i \le (n-1)} |n_{\mathrm{nz}}\, c_i|^{1/(n-i)}.$$

The test problem was the Chebyshev integration problem, in which the polynomial coefficients with $a_n = 1$ were computed as follows:

$$\begin{cases} a_{n-(2k-1)} := 0 \\ a_{n-2k} := -\sum_{2j+1}^{k} a_{n-2(k-j)} \end{cases} \tag{6}$$

where $k = 1, 2, \ldots, \lfloor n/2 \rfloor$.

Table 6 presents the benchmark results obtained by applying the DK method with DD, TD, and QD precision arithmetic to the Chebyshev integration problem, as described previously.

Table 6. Benchmark test (unit: second) for the Chebyshev integration problem ($n = 64$)

Snapdragon	DD	TD	QD	EPYC	DD	TD	QD
NoBF	0.0528	0.174	0.59	NoBF	0.0627	0.306	0.989
NoBF Neon	0.0381	0.172	0.47	NoBF AVX2	0.0431	0.251	0.608
BF	—	0.14	0.383	BF	—	0.214	0.628
BF Neon	—	0.111	0.336	BF AVX2	—	0.156	0.415

Table 6 lists the benchmark results for the Chebyshev integration problem ($n = 64$). The results with and without SIMD vectorization—Neon on the Snapdragon platform and AVX2 on the EPYC platform—are listed side-by-side (NoBF Neon/AVX2 and BF Neon/AVX2), and the effects of BF conversion (BF) are compared for TD and QD precisions. The effect of BF conversion is pronounced for TD and QD; on the EPYC, BF achieves an approximately 1.43× speedup for TD and 1.57× for QD relative to NoBF, and even when combined with SIMD (NoBF AVX2→BF AVX2), speedups of approximately 1.61× for TD and 1.47× for QD are obtained. We can observe similar trends for the Snapdragon, where the combination of BF conversion and Neon vectorization is effective for TD and QD; in particular, TD achieves an approximately 1.55× speedup in the NoBF Neon→BF Neon transition. These results confirm that the combined use of BF conversion and SIMD vectorization is effective for multi-component multiple-precision arithmetic with TD and QD precisions.

5 Conclusion and Future Work

In this study, we applied BF algorithms that suppress conditional branching to multicomponent multiple-precision arithmetic (DD/TD/QD) and evaluated their performance on CPU SIMD (AVX2 on EPYC) and Arm Neon. Benchmark tests were conducted for real and complex square matrix multiplications, real- and complex-valued evaluations of real-coefficient polynomials, and the Chebyshev integration problem. The results confirm that, for high precision (particularly TD and QD), the combined use of BF conversion and SIMD vectorization is effective and reduces the computation time compared with the standard implementation (NoBF). In addition, the results validate that BF conversion offers no benefit to DD precision arithmetic. These findings indicate that achieving performance improvements in multicomponent arithmetic requires (i) branch elimination (BF conversion) and (ii) the exploitation of instruction-level parallelism (SIMD vectorization) to be combined with an implementation strategy selected according to the target problem, precision level, and architecture.

The following directions are identified for future research:

1. **Adaptive implementation selection (autotuning)**: As the optimal execution path (NoBF/BF, SIMD width, and loop partitioning) varies with the precision level (DD/TD/QD) and problem size, we will introduce a mechanism for selecting the optimal implementation at either runtime or build time.
2. **Support for wider SIMD extensions and instruction sets**: Extensions to AVX-512 and Arm SVE/SVE2 will be pursued, along with establishing design guidelines for data layout and reduction operations as register width increases.
3. **Memory access optimization and enhanced parallelization**: In addition to improving cache efficiency in matrix operations and polynomial evaluation, we will optimize the coordination of thread-level parallelism via OpenMP with SIMD vectorization, including NUMA placement and scheduling.
4. **Extension to application problems**: The scope of the application will be expanded to include high-precision linear algebra problems (QR decomposition, eigenvalue problems, and iterative methods), special function evaluation, numerical integration, and optimization, and the effectiveness in practical settings (convergence behavior, total runtime, and energy efficiency) will be evaluated.

Through these efforts, we aim to realize accelerated multicomponent high-precision arithmetic in a more general and portable form by establishing a high-precision numerical computing infrastructure applicable to a wide range of computational environments, ranging from mobile devices to supercomputers.

Acknowledgment. This research was supported by JSPS KAKENHI (grant number 23K11127).

References

1. Bailey, D.H.: QD. https://www.davidhbailey.com/dhbsoftware/
2. Dekker, T.J.: A floating-point technique for extending the available precision. Numer. Math. **18**(3), 224–242 (1971)
3. Enge, A., Théveny, P., Zimmermann, P.: MPC. http://www.multiprecision.org/mpc/
4. Fabiano, N., Muller, J.M., Picot, J.: Algorithms for triple-word arithmetic. IEEE Trans. Comput. **68**, 1573–1583 (2019)
5. Granlaud, T., GMP development team. The GNU Multiple Precision arithmetic library. https://gmplib.org/
6. Hishinuma, T., Fujii, A., Tanaka, T., Hasegawa, H.: AVX acceleration of DD arithmetic between a sparse matrix and vector. In: Wyrzykowski, R., Dongarra, J., Karczewski, K., Waś Niewski, J., eds., Parallel Processing and Applied Mathematics, pp. 622–631, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg (2014)
7. Kotakemori, T., Fujii, S., Hasegawa, H., Nishida, A.: Lis: Library of iterative solvers for linear systems. https://www.ssisc.org/lis/
8. Kouya, T.: BNCmatmul. https://github.com/tkouya/bncmatmul
9. Kouya, T.: Acceleration of multiple precision matrix multiplication based on multi-component floating-point arithmetic using avx2. In: Gervasi, O., et al., eds., Computational Science and Its Applications – ICCSA 2021, pp. 202–217, Cham (2021). Springer International Publishing
10. Kouya, T.: Performance evaluation ofÂ accelerated complex multiple-precision lu decomposition. In: Gervasi, O., et al., eds., Computational Science and Its Applications, ICCSA 2024 Workshops, pp. 3–19, Cham (2024). Springer Nature, Switzerland
11. Lange, M., Rump, S.M.: Faithfully rounded floating-point computations. ACM Trans. Math. Softw. **46**, No. 3 (July 2020)
12. Lu, M., He, B., Luo, Q.: Supporting extended precision on graphics processors. In: Proceedings of the Sixth International Workshop on Data Management on New Hardware, DaMoN '10, pp. 19–26, New York, NY, USA, (2010). ACM
13. MPLAPACK/MPBLAS. Multiple precision arithmetic LAPACK and BLAS. https://github.com/nakatamaho/mplapack
14. Mukunoki, D., Ozaki, K.: Sparse iterative solvers using high-precision arithmetic with quasi multi-word algorithms. In: 2025 IEEE 18th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC), pp. 33–40. (2025)
15. MPFR Project. The MPFR library. https://www.mpfr.org/
16. Zhang, D.K., Aiken, A.: Automatic verification ofÂ floating-point accumulation networks. In: Piskac, R., Rakamarić, Z., eds. Computer Aided Verification, pp. 215–237, Cham 2025. Springer Nature, Switzerland (2025)
17. Zhang, D.K., Aiken, A.: High-performance branch-free algorithms for extended-precision floating-point arithmetic. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC '25, p. 695–710, New York, NY, USA, 2025. Association for Computing Machinery (2025)
18. Kouya, T.: Trial Approach to Accelerate Multi-component-type Multiple-precision Basic Linear Computation with Arm Neon Intrinsics (in Japanese). Technical report, HPC (2025)

Advection of Smooth and Non-smooth Tracers Using Spectral Bicubic Interpolation

Takeshi Enomoto(✉)

Disaster Prevention Research Institute, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
enomoto.takeshi.3n@kyoto-u.ac.jp
https://www.dpac.dpri.kyoto-u.ac.jp/

Abstract. This paper validates a semi-Lagrangian model employing the spectral bicubic interpolation scheme against a challenging suite of two-dimensional tracer advection test cases. Unlike traditional methods, the proposed scheme utilizes derivatives calculated in spectral space rather than physical space. The proposed approach is compared against three alternatives: an Eulerian model, a semi-Lagrangian model with quasi-cubic interpolation, and a non-interpolating semi-Lagrangian model.

The proposed spectral bicubic scheme consistently outperforms the alternative methods across several metrics. Advection tests using twin Gaussian hills demonstrate that the spectral bicubic scheme on a linear grid achieves the superior error convergence rate. Furthermore, the spectral bicubic scheme benefits significantly from relaxed truncation wave numbers, an advantage not observed with the quasi-cubic scheme. In simulations involving non-smooth tracer distributions, such as twin cosine bells and slotted cylinders, the spectral bicubic method exhibits fewer Gibbs ripples than the quasi-cubic approach.

The scheme also demonstrates superior mixing properties, as evidenced by the correlation between the advection of twin cosine and correlated cosine bells. While the unlimited spectral bicubic scheme does not inherently eliminate Gibbs ripples, this noise can be substantially mitigated using a quasi-monotonic short-wave filter. Finally, the model can be configured for global mass conservation with negligible computational overhead and only a minimal impact on overall accuracy.

Keywords: Numerical weather prediction · Tracer advection · Semi-Lagrangian method

1 Introduction

Advection is a fundamental process in fluid dynamics, playing a central role in the nonlinear regime. In the atmosphere, constituents such as water vapor and ozone are primary sources of the diabatic heating that drives the tropospheric

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 525–541, 2027.
https://doi.org/10.1007/978-3-032-30521-3_33

and stratospheric circulation. Consequently, transport modeling remained a focal point for researchers, leading to numerous state-of-the-art transport models [6] designed to achieve desirable properties such as high accuracy, monotonicity, shape preservation [12].

The first semi-Lagrangian (SL) advection model on the sphere, which update the tracer values by interpolating at the departure points of backward-in-time trajectories, was validated against an Eulerian model using a smooth Gaussian hill in solid-body rotation over the North Pole [14]. Unlike the Eulerian model, which exhibited downstream dispersion ripples, the SL model remained stable and accurate. However, standard SL models often employ cubic interpolation, which is prone to numerical dissipation. To address this, the non-interpolating SL model (NISL) [13] reduces numerical dissipation by decomposing advection into the approximate Lagrangian component, derived from the nearest grid point, and a residual Eulerian component.

Quasi-monotonicity in numerical weather prediction (NWP) is frequently achieved using a simple limiters that restrict the interpolated values to the range of surrounding grid points [2,21]. However, this approach can induce damping by clipping the local extrema. To mitigate this, the filters may be excluded at non-monotonic intervals [19], while more sophisticated selection methods can further reduce Gibbs ripples [8,24] (See Subsect. 2.5 for details).

Because the advective-form SL scheme is not formally conservative, *a posteriori* corrections, or "fixers," are often employed to restore mass balance in NWP and climate models. This study utilizes a lightweight mass correction scheme based on a variational formulation [18], which imposes minimal changes to the solution to restore the global mass without degrading accuracy. In contrast, flux-form SL schemes [7], based on the upwind-biased finite-volume scheme [22], are inherently monotonic and formally conservative. These schemes also preserve correlations among different tracers, benefiting chemical transport models (CTMs) [16,17].

In models using the spectral transform method [9], advective terms are typically calculated in physical space. While Eulerian models benefit from accurate spectral derivatives, most SL models relinquish this advantage. To reintroduce spectral accuracy into the advective-form SL framework, the spectral bicubic interpolation (SBC) scheme was proposed [3], utilizing derivatives calculated in spectral space. Previous tests with Gaussian hill advection demonstrated that SBC outperforms traditional interpolation by simultaneously eliminating dispersion and reducing numerical dissipation. Furthermore, SBC yields global mass errors more than three order of magnitude smaller than other schemes. SBC has since been extended to a non-negative filter via arcsine transformation with Hermine cubic interpolation [4].

The purpose of this study is to validate SBC using more challenging tracer distributions in deformational flows and to evaluate its performance in both unlimited and limited cases [5]. Comparisons are made among the SL model with SBC, the Eulerian model (EUL), the SL model with the quasi-cubic interpolation (QC) scheme and non-interpolating SL model (NISL).

The remainder of the paper is organized as follows. Section 2 describes the two-dimensional advection models; Sect. 3 summarizes the test case suite; Sects. 4.1, 4.2, and 4.3 compare the unlimited models regarding convergence, overshooting, and mixing properties; Sect. 4.4 discusses the influence of quasi-monotonic filters and the global mass fixer; and Sect. 5 provides a summary and concluding discussion.

2 Two-Dimensional Advection Models on the Gaussian Grid

The study employs the models formulated in the previous work [3,14], updated with a two-time-level scheme [1] in place of the three-time-level scheme in the SL model. Preliminary tests with the rotation of a Gaussian hill demonstrate that the results from the two-time-level model are identical to those from the three-time-level model. Advecting winds at the midpoint are specified analytically rather than extrapolated in time to eliminate extrapolation error.

2.1 Governing Equation

The governing equation for two-dimensional advection of tracer q on the unit sphere is

$$\frac{\mathrm{d}q}{\mathrm{d}t} = 0 \tag{1}$$

where

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\partial}{\partial t} + \boldsymbol{v} \cdot \nabla \tag{2}$$

is the total derivative and $\boldsymbol{v} = (u, v)$ is the wind velocity, with u and v representing the zonal and meridional components, respectively.

2.2 Eulerian Model

The Eulerian model (EUL) integrates:

$$q(t + \Delta t) = q(t - \Delta t) - \frac{2\Delta t}{\cos^2 \theta} \left(U \frac{\partial q}{\partial \lambda} + V \cos\theta \frac{\partial q}{\partial \theta} \right) \tag{3}$$

where t is the present time, Δt is the time step, (λ, θ) is the longitude and the latitude, and $(U, V) = (u, v) \cos\theta$ is the wind image. The variable without a time level is evaluated at t. All the variables and their derivatives are evaluated at the grid points without staggering. The horizontal derivatives are evaluated in spectral space.

2.3 Interpolating Semi-Lagrangian Model

In the interpolating SL model, the values at the next time level are predicted simply via:

$$q(\boldsymbol{g}, t + \Delta t) = q(\boldsymbol{r}, t). \tag{4}$$

In the backward trajectory scheme, the arrival point $\boldsymbol{g}$ is on the grid, while the departure point $\boldsymbol{r}$ generally is not. The trajectory is calculated along the great circle [14]. Advecting winds at $t + 1/2\Delta t$ are obtained analytically from the prescribed flow defined in Sect. 3.

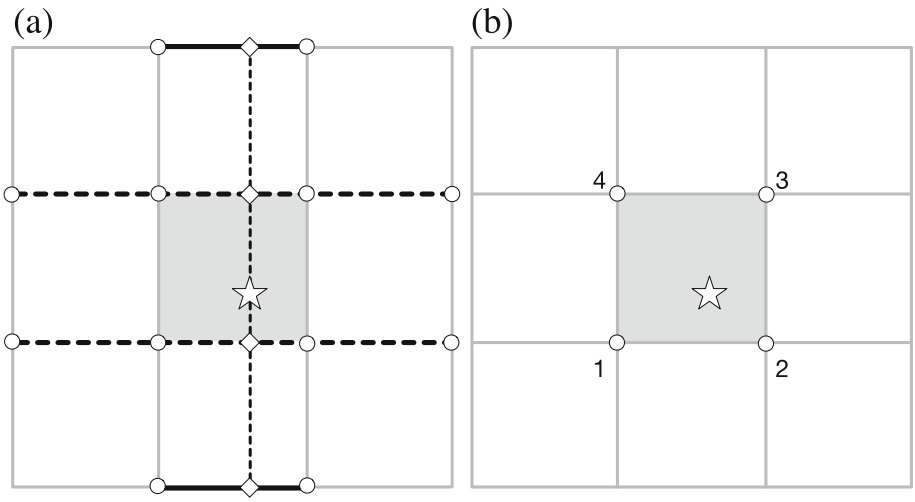


Fig. 1. Schematic diagrams for (a) the quasi-cubic interpolation and (b) the bicubic interpolation. The gray lines represent the model mesh. The value at the star located in the shaded stencil is found via interpolation using information on the surrounding grid points (open circles). The diamonds in panel (a) represent the temporary interpolation points at the same longitude as the interpolation point. Linear interpolation is used along the solid black lines, and the cubic Lagrange interpolation is used along the dashed black lines in panel (a). The grid-points of the stencil are labeled counter-clockwise from the bottom-left. (Color figure online)

Quasi-Cubic Interpolation. The QC scheme reduces the computational cost with minimal accuracy degradation compared to bicubic interpolation using the following steps [15, Fig. 1a].

1. Locate the stencil in which the target point (λ, θ) resides (denoted by a star).
2. Compute values at λ with one-dimensional linear interpolation along the two latitude circles an interval away from the stencil.
3. Compute values at λ with one-dimensional cubic Lagrange interpolation along the two latitude circles of the stencil.

4. Determine the value at (λ, θ) with one-dimensional cubic Lagrange interpolation along λ.

QC requires information from 12 grid points instead of 16 needed for bicubic Hermite or Lagrange interpolation.

Spectral Bicubic Interpolation. Bicubic interpolation is defined as

$$q = \sum_{i=1}^{4} \sum_{j=1}^{4} a_{ij} \alpha^{i-1} \beta^{j-1} \tag{5}$$

where α and β are the normalized local coordinates of the target point (denoted by the star in Fig. 1b) relative to node 1:

$$\alpha = \frac{\lambda - \lambda_1}{\Delta\lambda} \tag{6}$$

$$\beta = \frac{\theta - \theta_1}{\Delta\theta} \tag{7}$$

and (λ_1, θ_1) is the bottom-left corner of the stencil, from which the grid points are labeled counter-clockwise, and $\Delta\lambda, \Delta\theta$ are the grid intervals. The coefficients a_{ij} are determined from the values and the zonal, meridional, and cross derivatives [11]. Bicubic interpolation requires 16 values, but avoids the need for a halo when decomposed for parallel computation (Fig. 1b).

In the spectral bicubic scheme (SBC), derivatives are computed in spectral space:

$$\frac{\partial q}{\partial \lambda} = \mathrm{Re} \sum_m im q^m(\theta_j) \exp(im\lambda_i), \tag{8}$$

$$\frac{\partial q}{\partial \theta} = \mathrm{Re} \sum_m \sum_n q_n^m \exp(im\lambda_i) \frac{\mathrm{d}\tilde{P}_n^m(\theta_j)}{\mathrm{d}\theta}, \tag{9}$$

$$\frac{\partial^2 q}{\partial \lambda \partial \theta} = \mathrm{Re} \sum_m \sum_n im q_n^m \exp(im\lambda_i) \frac{\mathrm{d}\tilde{P}_n^m(\theta_j)}{\mathrm{d}\theta} \tag{10}$$

where $q^m(\theta_j)$ is the Fourier coefficient, q_n^m the spherical harmonic coefficient, and $\tilde{P}_n^m(\theta_j)$ the normalized associated Legendre function. The Fourier sum in m is typically computed via the Fast Fourier Transform (FFT). To save memory, the values of $\mathrm{d}\tilde{P}_n^m(\theta_j)/\mathrm{d}\theta$ are computed on the fly [20] using

$$\cos\theta_j \frac{\mathrm{d}\tilde{P}_n^m(\theta_j)}{\mathrm{d}\theta} = (n+1)\varepsilon_n^m \tilde{P}_{n-1}^m(\theta_j) - n\varepsilon_{n+1}^m \tilde{P}_{n+1}^m(\theta_j) \tag{11}$$

where

$$\varepsilon_n^m = \sqrt{\frac{n^2 - m^2}{4n^2 - 1}}. \tag{12}$$

2.4 Non-interpolating Semi-Lagrangian Model

The NISL model updates the values using:

$$q(\boldsymbol{g}, t + \Delta t) = q(\boldsymbol{r}^*, t - \Delta t) - 2\Delta t \boldsymbol{v}'(\boldsymbol{m}, t) \cdot \nabla q(\boldsymbol{m}, t) \tag{13}$$

where $\boldsymbol{r}^*$ is the grid point closest to the departure point, $\boldsymbol{v}' = \boldsymbol{v} - \boldsymbol{v}^*$ is the residual velocity, $\boldsymbol{v}^* = (\boldsymbol{g} - \boldsymbol{r}^*)/2\Delta t$ and $\boldsymbol{m}$ is the midpoint between $\boldsymbol{g}$ and $\boldsymbol{r}^*$ along a great circle trajectory. Interpolation is avoided for Lagrangian advection from the departure point, occurring only for the residual Eulerian advection at the midpoint. When the wind is weak enough that $\boldsymbol{g} = \boldsymbol{m} = \boldsymbol{r} = \boldsymbol{r}^*$, the model reduces to the Eulerian form (3). The phase speed error for the solid body rotation is proportional to $(m\omega'\Delta t)^2$; this is lower than the Eulerian error since $\omega' \leq \omega$ [14].

2.5 Shape-Preservation

Two selective filters are employed in shape-preserving cases for interpolating SL models.

The filter of SYS96 [19] suppresses the quasi-monotonic filter [2] in the interval $[x_i, x_{i+1}]$ if:

$$[f(x_i) - f(x_{i-1})][f(x_{i+2}) - f(x_{i+1})] < 0. \tag{14}$$

Then the over- and under shoots are limited by the global minimum and maximum [10].

To eliminate two-mesh Gibbs noise remaining with SYS96, the filter of [8] (NCS99) adds additional two conditions:

$$[f(x_{i-1}) - f(x_{i-2})][f(x_i) - f(x_{i-1})] > 0, \tag{15}$$

$$[f(x_{i+2}) - f(x_{i+1})][f(x_{i+3}) - f(x_{i+2})] > 0. \tag{16}$$

These constitute a five-mesh monotonicity condition. The quasi-monotonic filter is applied unless all conditions are met.

As advective-form SL models are not inherently conservative, we apply a global mass fixer SS04 [18]. Because the SBC scheme is nearly conservative [3], this variational fixer is chosen to restore mass with minimal solution perturbation.

3 Smooth and Non-smooth Test Cases

The standard test case suite [5] is briefly described in this section. These test cases are idealized yet modeled based on poleward advection from the tropics due to disturbances (Lauritzen, *pers. comm.*). While no analytical solution exists for the intermediate states, the initial distribution serves as the ground truth because the prescribed wind field is reversed at the midpoint of the integration period.

3.1 Initial Tracer Distributions

Four smooth and non-smooth distributions are defined using two centers at $\lambda_1 = 150°, \lambda_2 = 210°$ along the equator ($\theta_1 = \theta_2 = 0°$).

Gaussian Hills. The Gaussian hills are defined by:

$$q(\lambda, \theta) = \sum_{i=1}^{2} q_{\max} \exp\{-2b[1 - d_i]\} \tag{17}$$

where $q_{\max} = 0.95$, $b = 5$ and

$$d_i = \cos\theta \cos(\lambda - \lambda_i). \tag{18}$$

Cosine Bells. The cosine bells are defined by:

$$q_i(\lambda, \theta) = \begin{cases} \frac{q_{\max}}{2}[1 + \cos(\pi r_i / r_0)] & r_i < r_0 \\ 0 & r_i \geq r_0 \end{cases} \tag{19}$$

$$q(\lambda, \theta) = b + c \sum_{i=1}^{2} q_i(\lambda, \theta) \tag{20}$$

where $q_{\max} = 1$, $r_0 = 0.5$, $b = 0.1$, $c = 0.9$ and $r_i = \cos^{-1}(d_i)$.

Slotted Cylinders. The slotted cylinders are defined as:

$$q(\lambda, \theta) = \begin{cases} c & \text{in the following regions} \\ b & \text{otherwise} \end{cases} \tag{21}$$

where $b = 0.1$ and $c = 1$. The regions where $q = c$ are defined by $r_0 = 0.5$ and the following conditions:

- $r_i \leq r_0$, $|\lambda - \lambda_i| \geq \frac{r_0}{6}$ for $i = 1, 2$
- $r_1 \leq r_0$, $|\lambda - \lambda_1| < \frac{r_0}{6}$ and $\theta < -\frac{5}{12} r_0$,
- $r_2 \leq r_0$, $|\lambda - \lambda_2| < \frac{r_0}{6}$ and $\theta > \frac{5}{12} r_0$.

Correlated Cosine Bells. The correlated cosine bells are defined by:

$$q(\lambda, \theta) = \alpha q_c^2 + \beta \tag{22}$$

where q_c designates the tracer distribution of the cosine bells Sect. 3.1, $\alpha = -0.8$ and $\beta = 0.9$.

3.2 Advecting Flows

Among the non-divergent and divergent deformational flows [5], this study focuses on the non-divergent wind field:

$$u(\lambda, \theta, t) = \frac{10}{T} \sin^2(\lambda') \sin(2\theta) \cos\left(\frac{\pi t}{T}\right) + \frac{2\pi}{T} \cos(\theta) \tag{23}$$

$$v(\lambda, \theta, t) = \frac{10}{T} \sin(2\lambda') \cos(\theta) \cos\left(\frac{\pi t}{T}\right) \tag{24}$$

where $\lambda' = \lambda - 2\pi t/T$ and $T = 5$. The time-dependent coefficient $\cos(\pi t/T)$ reverses the flow direction at $t = T/2$, returning the tracer to its initial position at $t = T$. The second term in (23) represents solid-body rotation, which translates the tracer distribution over the period T.

4 Results

Advection tests are conducted at horizontal resolutions of 3, 1.5, 0.75, 0.375 and 0.1875°. These correspond to triangular (zonal and total) truncation wave numbers T39, T79, T159, T319 and T639, respectively, on a quadratic grid, which oversamples grid points at approximately three times the truncation wave number in the zonal direction to avoid aliasing. The resolution of the interpolating SL model with QC is defined by the truncation wave number of the quadratic grid, though time integration is performed in physical space. For the interpolating SL model with SBC, simulations using a linear grid are also conducted. A linear grid uses grid points the approximately twice the truncation wave number in the zonal direction, exploiting the absence of the nonlinear advection term in the SL formulation. These simulations use truncation wave numbers of TL59, TL119, TL239, TL479 and TL959.

All of the models are integrated with Courant-Friedrichs-Lewy (CFL) number CFL = 1.0. The CFL represents the ratio of the advected distance to the grid interval; in EUL schemes, this value typically defines the threshold for numerical stability. The SL models (QC, SBC and NISL) are also tested with CFL = 5.2.

4.1 Convergence

We first examine model convergence on quadratic grids using smooth tracer distribution, the Gaussian hills. Figure 2 displays the convergence of the standard error norms, defined following [23], with the correspoinding convergence rates calculated via linear fit (Table 1). The norms are formulated as follows [23].

$$\ell_1 = \left[\frac{I(|q - q_t|)}{I(|q_t|)}\right], \tag{25}$$

$$\ell_2 = \left\{\frac{I[(q - q_t)^2]}{I(q_t^2)}\right\}^{1/2}, \tag{26}$$

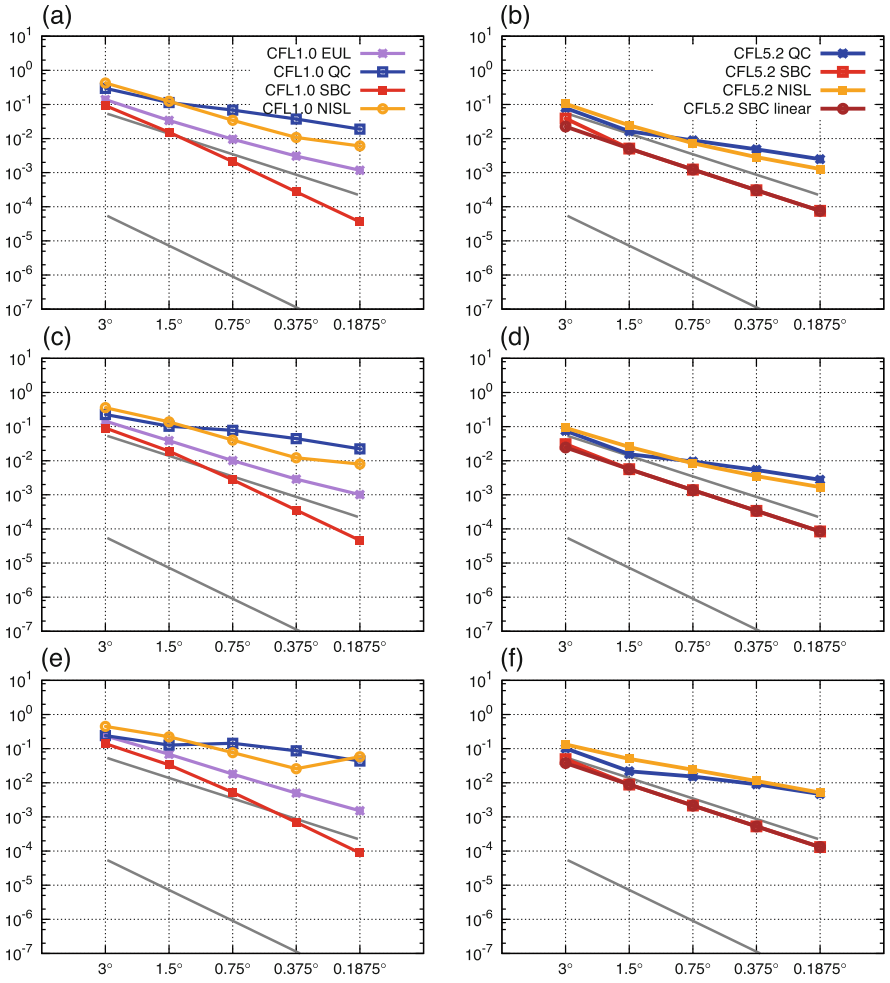


Fig. 2. Convergence of ℓ_1 (a, b), ℓ_2 (c, d) and ℓ_∞ (e, f) in the Eulerian model (EUL, purple), the SL models using the quasi-cubic interpolation (QC, blue) and using the spectral bicubic interpolation (SBC, red and brown for the quadratic and linear grids, respectively), and the non-interpolating SL model (NISL, orange) with the Gaussian hills for CFL = 1.0 (a, c ,e) and CFL = 5.2 (b, d, f). The upper and lower gray lines represent the second- and third-order convergences, respectively. (Color figure online)

$$\ell_\infty = \left[\frac{\max_{\forall\lambda,\theta} |q - q_\mathrm{t}|}{\max_{\forall\lambda,\theta} |q_\mathrm{t}|} \right], \tag{27}$$

$$q_\mathrm{max} = \frac{\max_{\forall\lambda,\theta}(q) - \max_{\forall\lambda,\theta}(q_\mathrm{t})}{\Delta q_0}, \tag{28}$$

$$q_\mathrm{min} = \frac{\min_{\forall\lambda,\theta}(q) - \min_{\forall\lambda,\theta}(q_\mathrm{t})}{\Delta q_0} \tag{29}$$

where q_t is the analytical value and

$$\Delta q_0 = \max_{\forall \lambda,\theta}[q(t=0)] - \min_{\forall \lambda \theta}[q(t=0)], \tag{30}$$

$$I(q) = \frac{1}{4\pi}\int_0^{2\pi}\int_{-\pi/2}^{\pi/2} q(\lambda,\theta,t)\cos\theta \mathrm{d}\lambda \mathrm{d}\theta. \tag{31}$$

EUL error norms follow a second-order convergence pattern, with rates of 1.7 for ℓ_1 and 1.8 for ℓ_2 and ℓ_∞ (Table 1). As traditional spectral models are Eulerian, these values serve as a reference.

QC and NISL exhibit degraded convergence rates. In contrast, SBC consistently outperforms the other schemes, achieving nearly third-order convergence (slopes in all norms ≈ 3.0, Table 1) with CFL = 1.0 (Fig. 2a, c, e) and second-order convergence with CFL = 5.2 (Fig. 2b, d, f). At low resolutions the SL models (QC, SBC, and NISL) yield smaller errors with CFL = 5.2, likely due to the reduced frequency of interpolations.

Utilizing a linear grid does not improve the convergence rate for Gaussian hills; rates are identical between quadratic and linear grids for QC and SBC with CFL = 1.0. At CFL = 5.2, the SBC convergence rate using the linear grid (brown) appears slightly lower than on the quadratic grid (red). This occurs because SBC significantly reduces errors at coarse resolutions; for example, at $\Delta\lambda = 3.0$, ℓ_1, ℓ_2 and ℓ_∞ values for SBC are substantially improved when moving from quadratic to linear grids.

Figure 3 illustrates the equatorial cross-sections of q for QC and SBC schemes. SBC (red) exhibits considerably less damping of the tracer amplitudes compared to QC (blue). When utilizing the linear grid, SBC (brown) retains the peaks ever more effectively. Furthermore, the numerical ripples observed between 45°E and 90°E are greatly reduced for the SBC scheme on the TL59 linear grid.

4.2 Non-smooth Distributions and Oscillations

Figure 4 illustrates the slotted cylinders at $t = T/2$ and $t = T$ for QC and SBC. Both schemes efficiently leverage the high resolution to maintain the tracer discontinuities. While the spatial extent of ripples in QC is limited, it exhibits significant over- and under-shooting near the cylinder edges (Fig. 5). SBC also produces oscillations, but with considerably smaller amplitudes.

Table 2 compares ℓ_2, $q_{\max}$ and $q_{\min}$ for QC and SBC. Across all resolutions, SBC maintains a smaller ℓ_2 norm. Notably, the absolute magnitudes of over- and under-shooting in QC increase with resolution, indicating that higher grid density exacerbates these numerical artifacts for the QC scheme.

4.3 Mixing Diagnostics

Mixing diagnostics evaluate the preservation of nonlinear relationship between two tracer distributions [5]. Figure 6 shows scatter plots between cosine bells (χ) and correlated cosine bells (ξ) for T159 and CFL = 1.0. The curve represents

Table 1. Convergence rates of the Eulerian model (EUL), the interpolating SL models using quasi-cubic interpolation (QC) and using spectral bicubic interpolation (SBC) and the non-interpolating SL model (NISL) with quadratic or linear grids for the Gaussian hills.

model	grid	CFL	ℓ_1	ℓ_2	ℓ_∞
EUL	quadratic	1.0	1.7	1.8	1.8
QC	quadratic	1.0	1.0	0.8	0.6
QC	linear	1.0	1.0	0.8	0.6
SBC	quadratic	1.0	2.8	2.8	2.7
SBC	linear	1.0	2.8	2.8	2.7
NISL	quadratic	1.0	1.6	1.4	0.9
QC	quadratic	5.2	1.2	1.1	1.0
QC	linear	5.2	1.2	1.1	1.0
SBC	quadratic	5.2	2.2	2.1	2.1
SBC	linear	5.2	2.0	2.0	2.0
NISL	quadratic	5.2	1.6	1.4	1.2

Table 2. Resolution dependency of ℓ_2 for the slotted cylinders, $q_{\max}$ and $q_{\min}$ for QC and SBC for CFL = 5.2 and a quadratic grid.

	ℓ_2		$q_{\max}$		$q_{\min}$	
$\Delta\lambda$	QC	SBC	QC	SBC	QC	SBC
3.	0.3437	0.3010	0.1464	0.1710	−0.8606	−0.1657
1.5	0.2631	0.2180	0.1874	0.1509	−0.1821	−0.1353
0.75	0.2002	0.1637	0.1904	0.1721	−0.1966	−0.1275
0.375	0.1534	0.1218	0.2361	0.1320	−0.2237	−0.1036
0.1875	0.1195	0.09314	0.3103	0.1131	−0.2842	−0.1104

the analytical relationship (22), with deviations categorized as "real" mixing (ℓ_r), unmixing (ℓ_u), or over-shooting (ℓ_o), defined following [5].

SBC has significantly smaller mixing in all norms. QC and NISL have a number of points that deviate from the curve.

SBC exhibits significantly less mixing across all norms. Conversely, QC and NISL show considerable deviations from the curve, characterized primarily by over-shooting and "real" mixing, respectively. EUL produces markedly lower unmixing and over-shooting errors than QC for CFL = 1.0.

4.4 Shape-Preserving Cases

Finally, we evaluate the impact of shape-preserving filters and a mass fixer. While such schemes can sometimes degrade convergence, the selective filters used here (Sect. 2.5) are designed to reduce noise while preserving local extrema. Figure 7

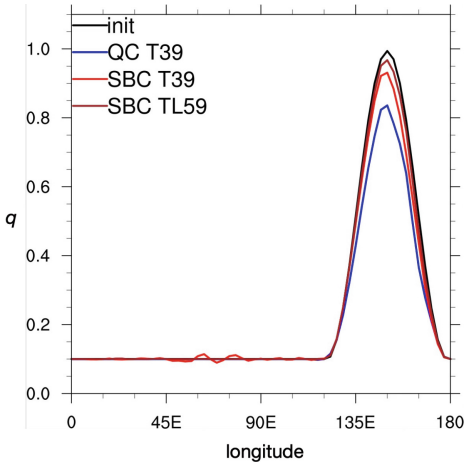


Fig. 3. Equatorial cross-sections of the cosine bells in longitudes between 0 and 180 at the initial time (black) at $t = T$ for QC T39 (blue), SBC T39 (red) or SBC TL59 (brown) and for CFL = 5.2. (Color figure online)

compares the equatorial cross-sections of the cosine bells at $t = T$ for SBC T39. The spurious ripples in the unlimited case (solid) are reduced by SYS96 (broken), but more eliminated by NCS99 (dotted). This suggests that the three-mesh condition (14) is insufficient to fully suppress high frequency Gibbs noise. The NCS99+SS04 case behaves similarly to NCS99, with only a slight reduction in peak amplitude due to the mass conservation constraint.

The shape-preserving filters effectively eliminate over- and under-shooting in the slotted cylinder test (Fig. 8). Both QC and SBC reproduce the tracer distribution well, removing artifacts near steep gradients. While QC retains the maximum value of 1.0 more effectively, SBC maintains a steeper gradient. Accuracy improves for QC in the shape-preserving case but decreases slightly for SBC (Table 3). Consequently, while both schemes perform similarly under these constraints, SBC retains marginal advantage across all resolutions.

5 Concluding Remarks

A two-dimensional advection model using the SL scheme with SBC has been compared against EUL, the SL model with QC and NISL in a suite of test cases [5]. In the convergence tests using twin Gaussian hills, SBC is found to produce the smallest errors and the best error-convergence rate. Furthermore, SBC benefits from the use of a linear grid with a larger truncation wave number for the same number of grid points, whereas QC does not. Cross-sections of the cosine bells and slotted cylinders show that the amplitudes of the over- and under- shooting in QC are large near the tracer boundaries, even if their spatial extent is limited. While SBC produces over- and under-shooting that spreads more extensively, the amplitudes are smaller than those of QC.

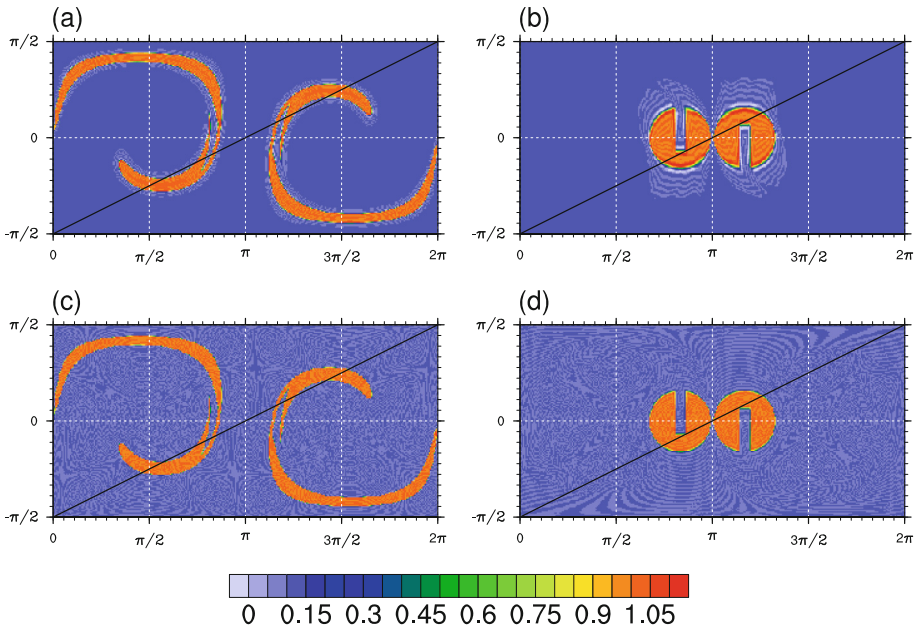


Fig. 4. Tracer distributions for QC (a, b) and SBC (c, d) with the slotted cylinders at $t = T/2$ (a, c) and $t = T$ (b, d) for CFL = 5.2 and for T639.

Table 3. Same as Table 2 but using the short-wave filter [8] and the global mass correction [18].

	ℓ_2		$q_{\max}$		$q_{\min}$	
$\Delta\lambda$	QC	SBC	QC	SBC	QC	SBC
3.	0.3623	0.3363	0.	0.	0.	0.
1.5	0.2794	0.2493	0.	0.	0.	0.
0.75	0.2120	0.1882	0.	0.	0.	0.
0.375	0.1612	0.1397	0.	0.	0.	0.
0.1875	0.1211	0.1048	0.	0.	0.	0.

Mixing diagnostics show the advantages of SBC; its mixing norms are smaller than those of alternative methods, and it benefits from relaxed truncation, particularly at coarse resolutions. Notably, over-shooting and "real" mixing are prominent in QC and NISL, respectively.

The computational tests demonstrate that SBC outperforms EUL, QC and NISL. SBC yields smaller error norms with ideal convergence and reduced over- and under-shooting, which can be further controlled using a selective filter and excellent mixing properties. While SBC requires an additional inverse Legendre transform to obtain meridional derivatives, the heuristic computational gain is

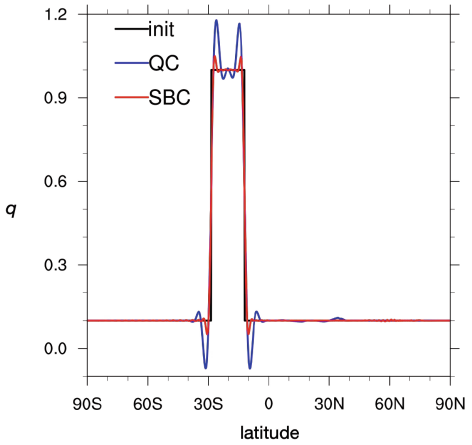


Fig. 5. Meridional cross-sections of q at 150°E at $t = 0$ (black) and at $t = T$ for QC (blue) and SBC (red) for T639 and for CFL = 5.2. (Color figure online)

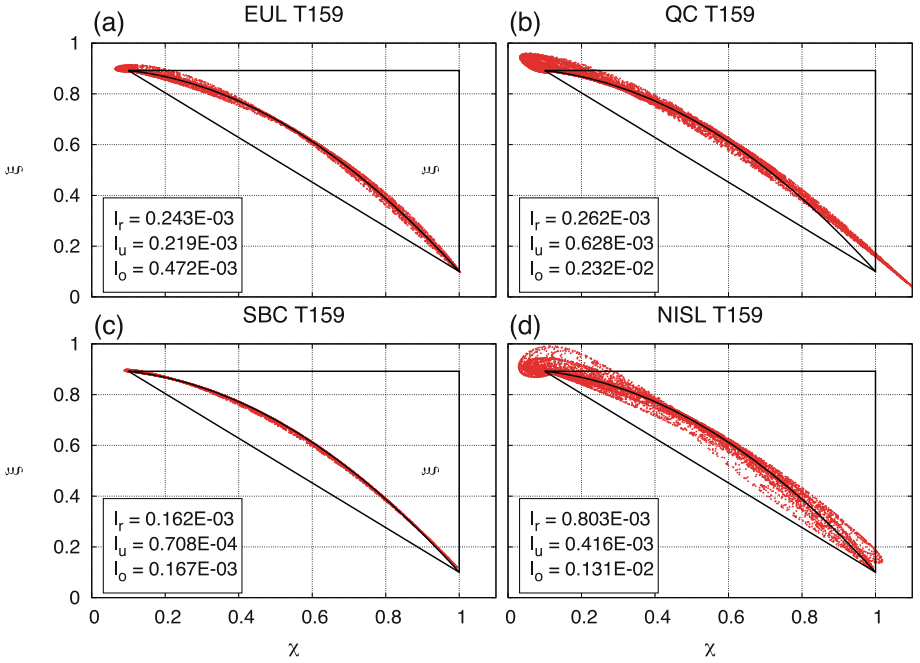


Fig. 6. Scatter plots displaying the relationship between the cosine bells and the correlated cosine bells for (a) EUL, (b) QC, (c) SBC and (d) NISL for CFL = 1.0 and T159 at $t = T/2$.

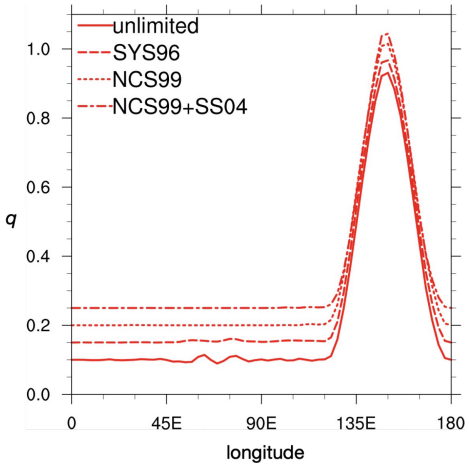


Fig. 7. Same as Fig. 3 with SBC T39 but using the short-wave filters and the global mass correction. Note that SYS96 (broken), NCS99 (dotted) and NCS99+SS04 (dash-dotted) are shifted upward by 0.05, 0.1 and 0.15, respectively, for clarity.

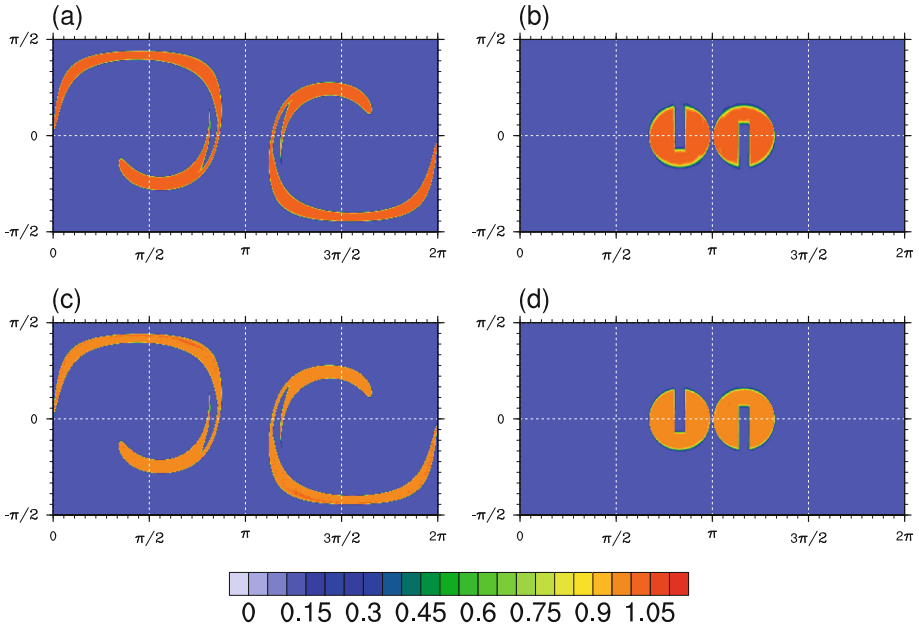


Fig. 8. Same as Fig. 4 but using NCS99 and SS04.

substantial. Because SBC achieves comparable accuracy to QC at half the truncation wave number, the total number of grid points is reduced by a factor of four. Furthermore, since the complexity of the spectral transforms scales with

the cube of the truncation wave number, the saving from operating at a coarser resolution significantly outweigh the overhead of the additional transform.

A somewhat unexpected finding is that the EUL can be more accurate than QC or NISL in mixing diagnostics. While SL advection schemes are often chosen over the Eulerian methods to avoid unrealistic ripples, these mixing diagnostics show that this is not always the case. However, if a larger time step (fewer interpolations) or a more accurate interpolation schem like SBC is used, SL mixing diagnostics improve significantly, outperforming EUL even for CFL = 1.0. These results underscore the importance of numerical accuracy alongside properties typically prioritized in tracer advection, such as conservation or monotonicity.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 15K13417, 21K03662, and 24H00021. The code is available from https://github.com/tenomoto/advection. The NCAR Command Language 6.3.0 (UCAR, 2015) and gnuplot 5.0 scripts provided by Dr Lauritzen were used for analysis and visualization with modification. The author thanks three anonymous reviewers for constructive comments.

Disclosure of Interests. The author does not have any competing interests.

References

1. Bates, J.R., McDonald, A.: Multiply-upstream, semi-Lagrangian advective scheme: analysis and application to a multi-level primitive equation model. Mon. Wea. Rev. **110**(12), 1831–1842 (1982). https://doi.org/10.1175/1520-0493(1982)110⟨1831:MUSLAS⟩2.0.CO;2
2. Bermejo, R., Staniforth, A.: The conversion of semi-Lagrangian advection schemes to quasi-monotone schemes. Mon. Wea. Rev., **120**(11), 2622–2632 (1992). https://doi.org/10.1175/1520-0493(1992)120⟨2622:TCOSLA⟩2.0.CO;2
3. Enomoto, T.: Bicubic interpolation with spectral derivatives. SOLA **4**, 5–8 (2008). https://doi.org/10.2151/sola.2008-002
4. Kashimura, H., Enomoto, T., Takahashi, Y.O.: Non-negative filter using arcsine transformation for tracer advection with semi-Lagrangian scheme. SOLA **9**, 125–128 (2013) https://doi.org/10.2151/sola.2013-028
5. Lauritzen, P.H., Skamarock, W.C., Prather, M.J., Taylor, M.A.: A standard test case suite for two-dimensional linear transport on the sphere. Geosci. Model Dev. **5**(3), 887–901 (2012). https://doi.org/10.5194/gmd-5-887-2012
6. Lauritzen, P.H., Coauthors: A standard test case suite for two-dimensional linear transport on the sphere: results from a collection of state-of-art schemes. Geosci. Model Dev., **7**(1), 105–145 (2014). https://doi.org/10.5194/gmd-7-105-2014
7. Lin, S.-J., Rood, R.B.: Multidimensional flux-form semi-Lagrangian transport schemes. Mon. Wea. Rev. **124**(9), 2046–2070 (1996). https://doi.org/10.1175/1520-0493(1996)124⟨2046:MFFSLT⟩2.0.CO;2
8. Nair, R., J. Côté, Staniforth, A.: Monotonic cascade interpolation for semi-Lagrangian advection. Quart. J. Roy. Meteor. Soc. **125**(553), 197–212 (1999). https://doi.org/10.1002/qj.49712555311
9. Orszag, S.A.: Transform method for the calculation of vector-coupled sums: application ot the spectral form of the vorticity equation. J. Atmos. Sci. **27**(6), 890–895 (1970). https://doi.org/10.1175/1520-0469(1970)027⟨0890:TMFTCO⟩2.0.CO;2

10. Ostiguy, L., Laprise, J.P.R.: On the positivity of mass in commonly used numerical transport schemes. Atmos.–Ocean, **28**(2), 147–161 (1990). https://doi.org/10.1080/07055900.1990.9649371
11. Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T.: Numerical recipes in Pascal. Cambridge University Press, Cambridge, UK (1989)
12. Rasch, P.J., Williamson, D.L.: Computational aspects of moisture transport in global models of the atmosphere. Quart. J. Roy. Meteor. Soc. **116**(495), 1071–1090 (1990). https://doi.org/10.1002/qj.49711649504
13. Ritchie, H.: Eliminating the interpolation associated with the semi-Lagrangian scheme. Mon. Wea. Rev. **114**(1), 135–146 (1986). https://doi.org/10.1175/1520-0493(1986)114⟨0135:ETIAWT⟩2.0.CO;2
14. Ritchie, H.: Semi-Lagrangian advection on a Gaussian grid. Mon. Wea. Rev. **115**(2), 608–619 (1987). https://doi.org/10.1175/1520-0493(1987)115⟨0608:SLAOAG⟩2.0.CO;2
15. Ritchie, H., Temperton, C., Simmons, A., Hortal, M. Davies, T., Dent D., Hamrud, M.: Implementation of the semi-Lagrangian method in a high-resolution version of the ECMWF forecast model. Mon. Wea. Rev. **123**(2), 489–514 (1995). https://doi.org/10.1175/1520-0493(1995)123⟨0489:IOTSLM⟩2.0.CO;2
16. Štajner, I., Riishøjgarrd, L.P., Rood, R.B.: The GEOS ozone data assimilation system: specification of error statistics. Quart. J. Roy. Meteor. Soc. **127**(573), 1069–1094 (2001). https://doi.org/10.1002/qj.49712757320
17. Sudo, K., Takahashi, M. Kurokawa, J. and Akimoto, H.: CHASER: a global chemical model of the troposphere 1. model description. J. Geophys. Res. **107**(4339), ACH 7-1–ACH 7–20 (2002). https://doi.org/10.1029/2001JD001113
18. Sun, W.-Y., Sun, M.-T.: Mass correction applied to semi-Lagrangian advection scheme. Mon. Wea. Rev. **132**(4), 975–984 (2004). https://doi.org/10.1175/1520-0493(2004)132⟨0975:MCATSA⟩2.0.CO;2
19. Sun, W.-Y., Yeh, K.-S., Sun, R.-Y.: A simple semi-Lagrangian scheme for advection equation. Quart. J. Roy. Meteor. Soc. **122**(533), 1211–1226 (1996). https://doi.org/10.1002/qj.49712253310
20. Temperton, C.: On scalar and vector transform methods for global spectral models. Mon. Wea. Rev. **119**(5), 1303–1307 (1991). https://doi.org/10.1175/1520-0493-119-5-1303.1
21. Temperton, C., Hortal, M., Simmons, A.: A two-time-level semi-Lagrangian global spectral model. Quart. J. Roy. Meteor. Soc. **127**(571), 111–127 (2001). https://doi.org/10.1002/qj.49712757107
22. van Leer, B.: Towards the ultimate conservative difference scheme. ii. Monotonicity and conservation combined in a second-order scheme. J. Comp. Phys. **14**(4), 361–370 (1974). https://doi.org/10.1016/0021-9991(74)90019-9
23. Williamson, D.L., Drake, J.B., Hack, J.J., Jakob, R., Swarztrauber, P.N.: A standard test set for numerical approximations to the shallow water equations in spherical geometry. J. Comp. Phys. **102**(1), 211–224 (1992). https://doi.org/10.1016/S0021-9991(05)80016-6
24. Zerroukat, M., Wood, N., Staniforth, A.: A monotonic and positive-definite filter for a semi-Lagrangian inherently conserving and efficient (slice) scheme. Quart. J. Roy. Meteor. Soc. **131**(611), 2923–2936 (2005). https://doi.org/10.1256/qj.04.97



Matrix Functions Computation in Hardware Accelerators

Daniel Oliveira[1(✉)], Juan Acebrón[2], and José Monteiro[3]

[1] INESC-ID, Instituto Superior Técnico - Universidade de Lisboa, Lisbon, Portugal
daniel.virtuoso.oliveira@tecnico.ulisboa.pt
[2] Universidad Carlos III de Madrid, Getafe, Spain
juacebro@math.uc3m.es
[3] INESC-ID, Instituto Superior Técnico - Universidade de Lisboa, Lisbon, Portugal
jcm@inesc-id.pt

Abstract. Solving very large linear systems or computing a function on a large matrix remains challenging due to the memory requirements and scalability limitations of classical numerical methods. Monte Carlo–based algorithms provide an attractive alternative because they naturally enable the computation of only the output elements of interest, significantly reducing memory pressure. In this work, a GPU-oriented Monte Carlo method was developed to solve the target problem directly while addressing the architectural challenges posed by modern GPUs. In particular, many Monte Carlo sampling procedures involve branching and irregular memory access patterns, which significantly degrade GPU performance. To mitigate these limitations, a novel data structure was introduced that combines ideas from Compressed Sparse Row (CSR) representations and Alias Tables. This structure enables efficient sampling while reducing memory overhead, thus improving parallelism on GPU architectures. The proposed approach was validated on real network matrices to evaluate both its accuracy and its scalability in practical large-scale settings.

Keywords: Matrix functions · Monte Carlo · GPU · WRS

1 Introduction

The efficient solution of very large linear systems and computing a function on a large matrix remain fundamental challenges in scientific computing. Classical numerical methods, particularly those based on direct factorizations or Krylov subspace iterations [1], are often limited by memory requirements, irregular data dependencies, and communication bottlenecks when matrices grow to tens or hundreds of millions of rows. In contrast, Monte Carlo–based linear algebra methods [2] provide a compelling alternative: they allow the computation of only the output elements of interest, avoid storing dense intermediate results, and exhibit natural parallelism due to the independence of their sampling procedures. Leveraging on these advantages, this work explores a stochastic approach

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 542–559, 2027.
https://doi.org/10.1007/978-3-032-30521-3_34

that directly targets the computation of matrix inverses and resolvent-based quantities while reducing global memory pressure.

At the core of this work is a Monte Carlo algorithm grounded on the probabilistic interpretation of matrix powers [3]. Instead of explicitly forming A^k, the algorithm estimates each desired entry via ensembles of random walks defined over the sparsity pattern of the matrix. This interpretation connected the method to a Taylor polynomial expansion [4] of a matrix function expressed as a finite sum of matrix powers, while avoiding explicit matrix-vector multiplications. Methods based on Krylov subspaces typically required repeated sparse matrix-vector multiplications, global reductions, and strict data-dependency patterns that limits their scalability. In contrast, the Monte Carlo formulation decomposes the computation into a very large number of short, independent random trajectories. This property proves particularly appealing for architectures with massive parallelism, where throughput is maximized by workloads that avoided synchronization and branching.

These characteristics make Graphics Processing Units (GPUs) a natural execution platform. Modern GPUs were designed for SIMD-style workloads with high arithmetic intensity and minimal control-flow divergence, and Monte Carlo sampling initially appeared well suited for such architectures due to its embarrassingly parallel nature. However, naive implementations suffer severe performance penalties. Many Monte Carlo steps require irregular memory accesses, stochastic branching, and row-dependent transition probabilities, all of which reduced SIMD efficiency and introduced warp divergence. Additionally, sampling from rows with widely varying degrees poses load-balancing challenges, and storing transition probabilities in formats such as Compressed Sparse Row (CSR) often result in costly random indexing. Consequently, the theoretical affinity between Monte Carlo algorithms and GPUs did not automatically translate into practical performance gains.

Building on these observations, this work proposes a GPU-oriented data structure that merges the compactness of CSR representations with the constant-time sampling capabilities of Alias Tables [5]. The resulting hybrid representation enables efficient Weighted Random Sampling (WRS) while significantly reducing warp divergence and irregular memory fetches. In addition, strategies for dynamic load balancing across GPU threads were investigated, which proved particularly important for matrices with highly skewed degree distributions. Potential cache-aware optimizations [6] were also explored as a means of further reducing sampling latency and improving overall throughput.

The methods developed in this work target applications in which selected entries of matrix functions encode meaningful structural and dynamical properties of large graphs. In network science, functions over the respective adjacency matrices have been widely used to quantify the importance of nodes and to characterize information flow in complex systems. Among these functions, the matrix exponential plays a central role because its entries naturally aggregate contributions from network walks of all lengths. This property gave rise to well-established measures such as communicability and subgraph centrality, which

have been used to study biological networks, social networks, and infrastructure systems [7]. For example, Estrada demonstrated that communicability derived from the matrix exponential provides a natural measure of how efficiently information propagates between nodes in complex networks.

Focusing on the matrix exponential therefore provided both a representative test case for the proposed Monte Carlo framework and a direct connection to interpretable application-level quantities. To evaluate the accuracy and scalability of the approach, the stochastic estimates were assessed using standard error metrics. In particular, we report the mean squared error and the maximum relative error between the Monte Carlo estimates and reference values, focusing on row-wise quantities of the matrix exponential. Experiments were conducted on synthetic graph models to enable controlled benchmarking across different sparsity patterns and degree distributions, followed by evaluations on large real-world networks to assess robustness and performance under realistic conditions.

2 Background

2.1 Matrix Functions

Matrix functions form a broad class of operations that extend scalar functions to matrices, including polynomials, rational functions, and analytic functions such as the exponential or the logarithm. They appear in many areas of numerical linear algebra and arise whenever a function defined on scalars is generalized to act on matrices through spectral or series-based constructions [8]. Efficient algorithms for evaluating matrix functions depend heavily on the structure of the matrix, such as sparsity, symmetry, or conditioning. In this work, matrix functions are considered through their Taylor polynomial expansions, which express the function as a finite (or truncated) sum of matrix powers; while this framework applies to a broad class of analytic functions, particular attention is devoted to the matrix exponential.

Polynomial series provide a fundamental framework for representing analytic functions as finite or infinite sums of powers. In particular, the Taylor series expresses a sufficiently smooth scalar function $f(x)$ as a polynomial expansion about a reference point x_0, with coefficients determined by the derivatives of the function at that point. Its general form is

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n, \tag{1}$$

assuming the series converges in a neighborhood of x_0. This construction extends naturally to matrix arguments, allowing matrix functions to be defined through convergent power series in the matrix variable. A canonical example is the matrix exponential, defined by

$$\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!}, \tag{2}$$

for a square matrix A. The factorial normalization ensures convergence of the series for all matrices and causes higher-order powers to decay rapidly, making polynomial truncations an effective representation of the matrix exponential.

2.2 Matrix Representation

In the context of complex networks, matrices are typically dominated by zero entries, making sparse storage essential. For such cases, the Compressed Sparse Row (CSR) format is particularly well suited, as it stores only non-zero elements and thus provides an efficient and effective representation.

The CSR format represents a sparse matrix using three one-dimensional arrays. The first array stores all non-zero values in row-major order. The second array stores the column indices corresponding to each non-zero value. The third array, often called the row pointer array, indicates the starting position of each row within the value and column index arrays. Since the algorithm presented later processes the matrix by rows, this row-based structure is adopted.

2.3 Alias Table

The Alias Table is a probabilistic data structure designed to enable constant-time sampling from a discrete probability distribution. Originally introduced as the Walker alias method [5], it transforms an arbitrary distribution into two lookup tables that allow random sampling in $O(1)$ time.

Let W be the sum of all weights and N the number of outcomes. During the alias table construction we define a target value

$$target = \frac{W}{N}.$$

Each outcome ideally contributes exactly this amount of probability mass. Outcomes whose weight exceeds the target are referred to as heavy, while those with weight below the target are light.

The construction process repeatedly pairs a light outcome with a heavy one. For each pair, a table entry is created that assigns the light outcome its full weight, and fills the remaining probability mass (up to the target) with a reference to the heavy outcome. The weight of the heavy outcome is then reduced by the amount assigned to the table entry; if its remaining weight falls below the target, it is reclassified as a light outcome, otherwise it may be paired again as a heavy outcome. This process continues until all table entries contain exactly the target probability mass.

As a consequence of the construction, each alias table entry contains probability mass from at most two outcomes, the bucket's primary outcome and a single alias, so it is impossible for any slot to contain three or more outcomes [5] (Fig. 1).

As a result, sampling can be performed by selecting a table index uniformly at random and then choosing between the primary outcome and its alias according

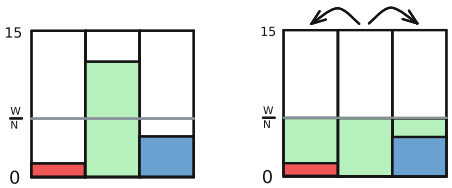


Fig. 1. Alias Table construction.

to the respective weights, using a single additional random draw. This described in Algorithm 1 and it can be observed that each sample is obtained in constant time.

Algorithm 1 Alias Method Sampling

```
i ← uniformly random integer from {0, ..., N − 1}
u ← uniformly random number in [0, 1]
if u < prob[i] then
    return i
else
    return alias[i]
```

Step-by-Step Example.

Consider the three weights , $w = (1, 11, 3)$. Then, $N = 3$, W $= 15$, $target = 5$. For each slot we store:

prob - how much of the slot is filled by the main outcome $\frac{w}{target}$ (a value in [0, 1]).
alias - the fallback outcome that fills the remaining portion of a slot beyond the main outcome.

For the example above, the alias table would be:

index	0	1	2
$prob[i]$	$1/5 = 0.2$	1	$3/5 = 0.6$
$alias[i]$	1	-	1

Interpretation:

- Index 0 contains 20% *prob*(ability) of itself (0) being the outcome and 80% of outcome being its *alias* (1)
- Index 1 contains 100% *prob*(ability) of itself (1) being the outcome

– Index 2 contains 60% *prob*(ability) of itself (2) being the outcome and 40% of outcome being its *alias* (1)

Since each slot is chosen with probability $\frac{1}{3}$, and the internal split reflects the original weights, the resulting samples follow the original distribution:

$$P(0) = \frac{1}{3} \cdot \frac{1}{5} = \frac{1}{15}, P(1) = \frac{1}{3} \cdot (\frac{4}{5} + \frac{5}{5} + \frac{2}{5}) = \frac{11}{15}, P(2) = \frac{1}{3} \cdot \frac{3}{5} = \frac{3}{15} \quad (3)$$

2.4 Network Metrics

A large number of metrics have been proposed to characterize complex networks, including degree-based measures, path-based distances, spectral indices, and centrality measures [9]. Many of these metrics are most naturally expressed in terms of matrix representations of networks, where connectivity patterns are encoded in an adjacency matrix. Among these, a particularly expressive family of metrics is derived from the matrix exponential. Because the matrix exponential naturally aggregates contributions from walks of all lengths, giving more weight to shorter walks, it provides a mathematically principled way to encode both local and global connectivity in a single operator. Several metrics can be derived from the matrix exponential, although only a subset is used in the present study.

The entries of the matrix exponential themselves have direct network interpretations. The diagonal entry $(e^A)_{ii}$ measures how strongly node i participates in closed walks, and can be interpreted as a form of subgraph centrality. Off-diagonal entries $(e^A)_{ij}$ quantify how strongly nodes i and j are connected through all possible paths in the network. Together, these entries provide fine-grained, node-level and pairwise information about network structure.

2.5 GPU

A Graphics Processing Unit (GPU) is a highly parallel processor composed of many lightweight cores designed to execute thousands of threads simultaneously. In contrast to CPUs, which optimize for low-latency execution of a few complex tasks, GPUs prioritize high throughput on large batches of similar operations.

GPUs follow a single-instruction-multiple-thread (SIMT) execution model, where groups of threads execute the same instruction concurrently on different data elements. Although each thread has its own program state, instructions are issued in lockstep for a group of threads, enabling efficient data-parallel execution. This execution model allows GPUs to exploit massive parallelism while maintaining a relatively simple control structure. GPUs typically share access to a unified global memory space, complemented by smaller, faster on-chip memories.

This architecture makes GPUs particularly effective for workloads that can be expressed as data-parallel computations, such as scientific simulations, machine learning, and image processing.

The parallel nature of GPU hardware strongly influences how algorithms must be designed and implemented. Threads on a GPU are organized into warps,

which are fixed-size groups of threads (typically 32 threads on NVIDIA GPUs) that execute instructions simultaneously in the SIMT model. Multiple warps are scheduled on a streaming multiprocessor (SM) to hide memory latency by switching execution when one warp stalls.

Because all threads within a warp execute the same instruction at a given time, branch divergence occurs when threads within a warp follow different execution paths due to conditional statements. In such cases, divergent paths are serialized, reducing effective parallelism and leading to performance loss. Similarly, inefficient memory access patterns, such as uncoalesced global memory accesses where threads in a warp access scattered memory locations, result in multiple memory transactions instead of a single efficient one, significantly increasing memory latency. Furthermore, insufficient parallel workload limits the GPU's ability to fully occupy its processing units, preventing effective latency hiding and under-utilizing available computational resources.

As a result, GPU performance depends not only on raw computational capability but also on how well memory accesses are organized, how control flow is aligned across warps, and how much parallelism can be exposed. GPU programming therefore requires careful consideration of the memory hierarchy, thread synchronization, and workload distribution to fully exploit the hardware.

3 Related Work

3.1 A Distributed Monte Carlo Based Linear Algebra Solver Applied to the Analysis of Large Complex Networks

Magalhães et al. [10] propose a distributed Monte Carlo (MC) framework for computing selected entries of matrix functions arising in large-scale network analysis, with a particular focus on matrix inverses and resolvent-based quantities. Their method builds on the classical probabilistic interpretation of matrix powers introduced by Ulam–von Neumann [3], where powers of a matrix are estimated through ensembles of random walks. By truncating the Neumann series expansion and estimating each term stochastically, the approach avoids explicit dense matrix formation and enables the computation of specific rows or entries. This selective computation is particularly well suited for network metrics such as Katz centrality and the resolvent Estrada index.

A key contribution of their work is a distributed-memory parallel implementation. The sparse matrix is partitioned row-wise across multiple compute nodes, and random walks migrate between nodes via message passing. While the independence of walks provides excellent parallel scalability, this model introduces communication overhead due to walk migration. To mitigate this, the authors propose buffering strategies, asynchronous communication, and workload-aware data partitioning. The resulting MPI/OpenMP implementation demonstrates strong scalability on large supercomputing systems, handling matrices with tens of millions of rows.

In addition, the authors introduce several optimizations, including reuse of partial random walks, simultaneous computation of multiple right-hand sides,

and memory-efficient formulations tailored to network metrics. Their results show that, although deterministic Krylov-based methods remain preferable for single linear solves, Monte Carlo methods can outperform classical approaches when computing traces, multiple inverse-vector products, or resolvent-based quantities at scale.

The present work adopts the same Monte Carlo random-walk framework, where matrix powers are estimated via Markov chains over the sparsity pattern and combined through a series expansion. The main algorithmic difference lies in the choice of expansion: while Magalhães et al. rely on Neumann (resolvent-based) series, this work targets the Taylor expansion of the matrix exponential, leading to different weighting of sampled paths.

More importantly, the two approaches differ in their computational model. Magalhães et al. focus on distributed-memory systems, where data is explicitly partitioned across nodes and communication is required to propagate random walks. In contrast, this work targets shared-memory architectures, particularly GPUs, where all data resides in a common memory space and parallelism is achieved through massive thread-level concurrency. As a result, the primary challenges shift from communication overhead to efficient memory access, load balancing, and minimizing warp divergence during sampling.

Despite these differences, several insights from the distributed setting remain relevant. In particular, strategies for managing large-scale data and irregular workloads can inform out-of-core GPU implementations, where data movement between host and device memory plays a role analogous to communication in distributed systems.

3.2 Skywalker: Efficient Alias-Method-Based Graph Sampling and Random Walk on GPUs

Wang et al. introduce Skywalker [11], a high-throughput GPU framework for large-scale graph sampling and random-walk generation. Random walks are a fundamental primitive in many graph analytics and representation learning algorithms, yet implementing them efficiently on GPUs is challenging due to irregular memory access patterns, load imbalance, and the cost of sampling from high-degree vertices. Skywalker addresses these challenges through a carefully designed system that combines alias-method-based sampling, GPU-oriented data structures, and dynamic load-balancing mechanisms to fully exploit GPU parallelism.

A key innovation of the framework is the use of the alias method for constant-time sampling of discrete probability distributions associated with graph transitions. While the alias method is widely used in probabilistic sampling, its efficient implementation on GPUs is nontrivial because constructing and accessing alias tables can incur substantial memory overhead and irregular memory accesses. Skywalker proposes a GPU-friendly alias-table representation and an optimized construction procedure that significantly reduces both preprocessing time and memory usage. This enables efficient sampling even for large graphs with highly

skewed degree distributions, where naive approaches would lead to severe performance degradation.

In addition to improving sampling efficiency, the system introduces several GPU-specific optimizations to mitigate load imbalance during random-walk execution. Because walks may vary in length and may encounter vertices with widely differing degrees, threads can experience significant divergence. Skywalker addresses this issue through dynamic workload management and warp-aware scheduling techniques that redistribute work among GPU threads to maintain high utilization. Together with memory optimizations for storing graph structures and sampling tables, these techniques enable the framework to achieve substantial performance improvements over prior CPU and GPU implementations of graph sampling workloads.

The present work shares several conceptual and algorithmic similarities with the Skywalker framework. Both approaches rely on random walks over sparse graph structures and must address similar challenges related to irregular memory access, stochastic branching, and GPU thread divergence. In particular, Skywalker's use of alias-based sampling directly relates to the hybrid CSR-alias data structures explored in this work for accelerating transition sampling in Monte Carlo estimators of matrix functions. However, the computational objectives of the two approaches differ significantly. Skywalker is designed as a general-purpose graph sampling engine supporting applications such as representation learning, whereas our work focuses on Monte Carlo estimators for matrix functions, particularly the matrix exponential, where random walks serve as a mechanism for approximating matrix-power expansions. Consequently, while Skywalker emphasizes throughput for large numbers of independent walks, the present work integrates similar sampling ideas within a stochastic linear-algebra framework aimed at estimating specific entries and traces of matrix functions.

4 Monte Carlo Approximation of the Matrix Exponential

The algorithm adopted in this work is based on the Monte Carlo matrix power method described in Sect. 3.1 While the original formulation is primarily presented in the context of matrix inversion and resolvent-based metrics, the same principles naturally extend to the computation of the matrix exponential.

4.1 Matrix Normalization

In order to ensure numerical stability of the Monte Carlo estimator, the input matrix is scaled prior to the simulation. In particular, let

$$\alpha = \max_i \sum_j |A_{ij}| \tag{4}$$

denote the maximum absolute row sum of A. We consider the normalized matrix

$$\widetilde{A} = \frac{1}{\alpha} A. \tag{5}$$

This scaling ensures that the row sums of $\widetilde{A}$ are bounded by 1, which in turn guarantees that the multiplicative factors accumulated during the random walks (cf. Algorithm 3) remain bounded. In particular, the function GetRowWeight returns values in $[0, 1]$, preventing exponential growth of the walk contributions and reducing estimator variance.

The matrix exponential of the original matrix can then be recovered as

$$e^A = e^{\alpha \widetilde{A}} = \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} \widetilde{A}^k, \tag{6}$$

which corresponds to reweighting the contributions of each sampled path by the factor α^k. This normalization step is therefore essential for maintaining both numerical stability and efficient Monte Carlo estimation.

4.2 Random Walk Interpretation

Following the formulation of [12], each entry of A^k can be expressed as the expected value of a random walk of length k over the indices of the matrix.

A random walk starting at row i evolves according to transition probabilities proportional to the entries of the current row. The contribution of a single walk is represented by a multiplicative random variable whose expected value corresponds to an entry of the matrix power A^k. By averaging over p independent walks (referred to as *plays*), the expectation can be approximated by a Monte Carlo estimator whose statistical error decreases proportionally to $\mathcal{O}(p^{-1/2})$.

When computing the matrix exponential, each estimated matrix power is further scaled by the corresponding Taylor coefficient $1/k!$, ensuring consistency with the series expansion.

4.3 Algorithm Overview

The algorithm proceeds row-wise. For each row i, the approximation of the corresponding row of e^A is obtained by accumulating Monte Carlo estimates of the powers $A^k/k!$ for $k = 2, \ldots, n$.

Algorithm 2 Monte Carlo Exponential Matrix

Require: A (matrix), n (series length), p (number of plays)
Ensure: B (approximation of e^A)

```
function ExponentialMatrix(A, n, p)
    N ← size(A)
    for i = 0 to N − 1 do
        B_{i*} ← [0, ..., 0]
        for k = 2 to n do
            B_{i*} ← B_{i*} + CalculateRow(A, i, k, p)
    return B
```

Algorithm 3 Monte Carlo Row Power Computation

```
function CalculateRow(A, i, k, p)
    N ← size(A)
    r[0 : N − 1] ← [0, . . . , 0]
    for m = 1 to p do
        value ← 1
        currentRow ← i
        for l = 0 to k − 1 do
            value ← value × GetRowWeight(A, currentRow)
            selectedColumn ← RandomWeightedChoice(A, currentRow)
            currentRow ← selectedColumn
        r[currentRow] ← r[currentRow] + value
    for j = 0 to N − 1 do
        r[j] ← r[j]/p
        r[j] ← r[j]/k!
    return r
```

The core computation is performed by Algorithm 3, which estimates a single row of A^k using Monte Carlo random walks.

An inspection of Algorithm 3 reveals that the dominant computational cost arises from the repeated sampling operations(RandomWeightedChoice) required during the random walks. For each row i and each power k, the algorithm performs p independent walks of length k, resulting in $\mathcal{O}(p \cdot k)$ sampling steps per row.

4.4 Random Weighted Choice

Let $\mathcal{S} = \{s_1, s_2, \ldots, s_d\}$ be a finite set of outcomes associated with non-negative weights $w_1, w_2, \ldots, w_d$. The RandomWeightedChoice operation, as used in Algorithm 3, selects an element $s_i \in \mathcal{S}$ with probability

$$P(s_i) = \frac{w_i}{\sum_{j=1}^{d} w_j}. \tag{7}$$

A standard implementation of RandomWeightedChoice proceeds by first normalizing the weights and constructing the cumulative distribution function (CDF). More precisely, let

$$c_i = \sum_{j=1}^{i} \frac{w_j}{\sum_{k=1}^{d} w_k}, \tag{8}$$

with $c_d = 1$. Sampling is then performed by drawing a uniform random variable $u \sim \mathcal{U}(0, 1)$ and selecting the smallest index i such that $c_i \geq u$. This can be implemented efficiently using binary search over the cumulative array, resulting in a time complexity of $\mathcal{O}(\log d)$ per sample.

While this approach is simple and widely used, it introduces logarithmic overhead and requires accessing multiple memory locations during the search. In the context of Monte Carlo random walks, where this operation is executed repeatedly, such costs become a significant performance bottleneck.

The Alias Table method provides an alternative sampling strategy that enables exact RandomWeightedChoice in constant time $\mathcal{O}(1)$ after a linear preprocessing step. However, classical alias tables assume dense index sets and do not integrate naturally with sparse matrix representations such as CSR.

4.5 Implementation Details

While the previous subsection focused on the probabilistic formulation of the sampling procedures, the practical performance of the method depends strongly on how these operations are implemented on parallel hardware.

The base of our GPU implementation consists of a single kernel that is launched where each block is assigned a row. Within each block, a strided loop processes the plays, after which the result is multiplied by the vector and stored. Shared memory is used at the warp level to store partial results from each thread, followed by a warp-level reduction and a final reduction across warps to obtain the final value. As for the matrix information, it uses a read-only cache method to capitalize on this propriety that the matrix information has.

For the binary search, the values are normalized and accumulated per row, allowing sampling through a uniformly generated random number followed by a standard binary search over the cumulative row weights. For the alias version it is implemented as described in Sect. 5

5 CSR–Alias Sampling Structure

To address the sampling bottleneck described above, we propose a hybrid data structure that enables constant-time weighted sampling directly from sparse matrix rows stored in Compressed Sparse Row (CSR) format.

5.1 Hybrid CSRâĂŞAlias Representation

Our approach embeds alias tables directly into the CSR representation. The standard CSR arrays *p_ row* and *colIdx* are preserved unchanged and continue to encode the sparse adjacency structure.

Instead of storing raw edge weights in the values array, each row is preprocessed into an Alias Table consisting of the two auxiliary arrays described in Sect. 2.3:

prob: probability threshold for the primary outcome
alias: fallback index used when the threshold test fails

Each row therefore stores a lightweight alias table corresponding to its non-zero entries. Since a row with degree d requires $\mathcal{O}(d)$ storage, the overall memory complexity remains

$$\mathcal{O}(E + N), \tag{9}$$

where E denotes the number of non-zero entries and N the number of rows. This matches the asymptotic storage complexity of standard CSR while enabling constant-time weighted sampling.

5.2 Example

Consider the first row shown in Fig. 2 containing three non-zero weights $(1, 11, 3)$. During preprocessing, these weights are converted into the new CSR-Alias representation.

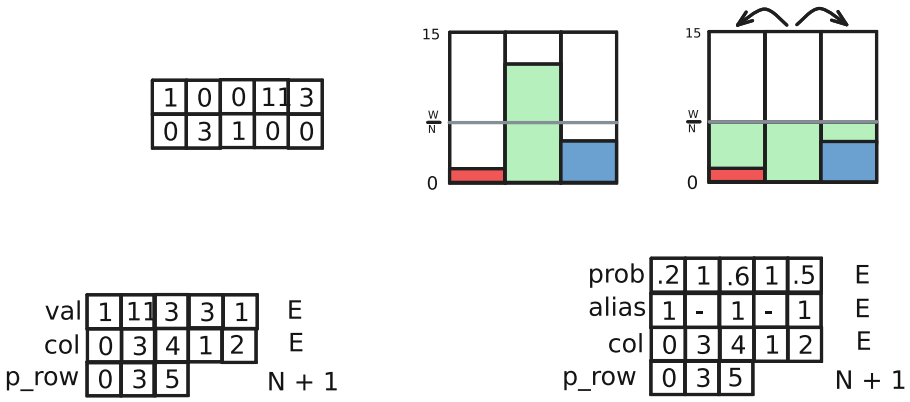


Fig. 2. Comparison between standard CSR representation (left) and the proposed CSR-Alias structure (right).

Each position in the table (right side) represents an equal share of the total probability mass. Sampling from this row proceeds as follows:

1. select an index uniformly at random,
2. compare a random number with the stored probability threshold,
3. redirect to the alias entry if necessary.

This procedure produces the correct weighted outcome in constant time.

5.3 Memory Overhead

An important aspect of the proposed representation is that it preserves the memory efficiency of the standard CSR format. For the example above, a standard

CSR2 (left side) representation of the row would store *rowPtr*, that encodes the starting and ending offsets of the row and the column and value arrays.

In the proposed CSR–Alias representation2 (right side), the structural components remain unchanged, and only the values array is replaced.

Thus, for each non-zero entry, the original scalar weight is replaced by two values: a probability threshold and an alias index. While this increases the per-entry storage by a constant factor, the overall memory complexity remains $\mathcal{O}(E)$, with E denoting the number of non-zero entries.

Importantly, the *rowPtr* and *colIdx* arrays are identical in both representations, ensuring that the sparse structure and traversal properties of CSR are fully preserved. In practice, the additional memory overhead is modest and is offset by the elimination of auxiliary sampling structures, such as cumulative distribution arrays, as well as by the improved runtime performance.

5.4 Parallel and GPU Suitability

The alias-based sampling procedure is particularly well suited for massively parallel architectures. Each sampling step consists of two random number generations, two memory accesses, and a single comparison. Unlike cumulative sum methods, the procedure avoids sequential memory access patterns and conditional branching, enabling warp-consistent execution on GPUs and improving parallel efficiency in large-scale stochastic simulations.

6 Results

This section evaluates the algorithm described in Sect. 4 in terms of numerical accuracy and computational performance on a GPU, highlighting the proposed CSR-Alias sampling structure. We first describe the execution environment, then analyze the estimator precision, and finally present runtime improvements.

6.1 Experimental Environment

All experiments were conducted on Deucalion, where the GPU is A100-40. The implementation was written in CUDA and compiled with `-O3` optimization.

For time measurements the value were taken from when the code starts building the Alias Table to when it already as the result calculated, but not printed. The Alias Table creation timer only represents a really small fraction of the time. But if for example we would be trying to get a low accuracy for a really big matrix this value could become a significant portion of the results. For the 131k nodes and 20 average degree the time to create the alias table was 30 ms. If this time becomes significant then Weighted Random Sampling on GPUs/citelehmann2021weighted offers some optimizations for this problem.

The Monte Carlo matrix exponential algorithm was executed on four sparse matrices. One matrix, `dictionary28`, was obtained from the SuiteSparse Matrix

Collection, while the remaining three were synthetically generated using graph models.

The *Watts–Strogatz* matrix was generated using the Watts–Strogatz model, consisting of approximately 65,000 nodes and 131,000 edges[1], with a rewiring probability of 0.8, resulting in a small-world structure with high randomness.

The matrix `kron_graph_u` was created using the Kronecker graph model, also with around 65,000 nodes and 171,000 edges[??], exhibiting a mostly uniform edge distribution.

Similarly, `kron_graph_c` was generated using the same Kronecker approach and shares the same size 65,000 nodes and 171,000 edges[??], but its nonzero entries are concentrated toward the top-left region of the matrix, reflecting a more localized structure .

The parameters used were:

- number of plays p (1M, 2M, 4M)
- maximum walk length n (30)

Unless otherwise stated, each configuration was executed multiple times and the average results are reported.

6.2 Estimator Precision

The Monte Carlo estimator used in this work follows standard convergence behavior, where the estimation error decreases proportionally to

$$\mathcal{O}(p^{-1/2}), \tag{10}$$

with p denoting the number of independent random walks. However, in our setting, the quantity of interest is not an individual entry of the matrix exponential, but rather the sum of each row of e^A. As if we were multiplying by a vector of 1. This corresponds to aggregating all estimated components of the row vector produced by the Monte Carlo procedure.

Since this is obtained as a sum of multiple Monte Carlo estimators, its variance benefits from aggregation effects. Assuming approximate independence between contributions, the variance of the sum decreases proportionally to the number of aggregated terms. In our case, each row accumulates approximately

$$N_{\text{agg}} \approx 6.5 \times 10^4 \tag{11}$$

contributions across different powers and random walks.

As a result, the overall estimation error follows

$$\mathcal{O}\left((p \cdot N_{\text{agg}})^{-1/2}\right), \tag{12}$$

[1] In the respective adjacency matrix representation, each node will have one corresponding row and column, and each undirected edge two non-zero entries in the matrix.

which provides a significant variance reduction compared to estimating individual entries.

Before our experiments, we calculated the error to be

$$\epsilon_i \leq \frac{1}{\sqrt{10^6 \times 10^5}} \tag{13}$$

To validate the accuracy of the estimator, reference values were computed using MATLAB's built-in matrix exponential function. In tests, the average relative error with respect to these reference values was

$$\bar{\epsilon} = 6.6 \times 10^{-6}. \tag{14}$$

These results confirm that the aggregation inherent in the row-wise summation significantly improves estimator stability, yielding accurate approximations even for moderate values of p.

6.3 Performance Evaluation

We compared the proposed CSR-Alias sampling method, as described in the algorithm in Sect. 4, against baseline implementations using binary normalized search. In all executions the relative error remained below 10^{-6} across all tested configurations.

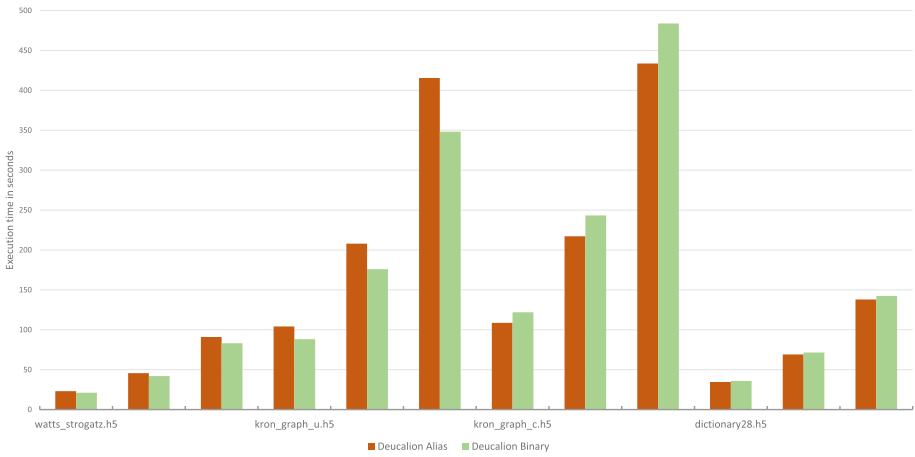


Fig. 3. Execution time comparison between the CSR–Alias implementation and the binary-search-based sampling across different datasets and configurations. Results are shown for varying numbers of plays (1M, 2M, 4M) on four graph instances: *watts_strogatz*, *kron_graph_u*, *kron_graph_c*, and *dictionary28*.

Figure 3 shows the total runtime of the Monte Carlo matrix exponential algorithm for different configurations.

For the first two instances, namely *Watts-Strogatz* and *Kronecker Uniform*, the baseline implementation based on binary search exhibits competitive performance. In these datasets, the majority of rows have very small degree (typically close to 2), which limits the depth of the binary search and reduces the number of branch operations. As a result, the baseline method benefits from low control-flow divergence and reduced memory overhead, leading to performance that is comparable to, and in some cases slightly better than, the alias-based approach.

In contrast, for the remaining instances characterized by higher average degree and more heterogeneous degree distributions, the advantages of the CSR-Alias method become significantly more pronounced. In these cases, the cost of binary search increases with the row size, both in terms of additional comparisons and irregular memory access patterns. The proposed method, by maintaining constant-time $\mathcal{O}(1)$ sampling, avoids this overhead and achieves substantially higher speedups.

Overall, these results highlight that the benefits of the CSR-Alias structure are particularly significant for graphs with moderate to high degrees, where weighted sampling constitutes a dominant portion of the computational cost.

6.4 Discussion of Speedups

The observed performance improvements can be attributed to the reduction in sampling complexity. While binary search sampling requires $\mathcal{O}(\log d)$ time per step, the proposed method achieves constant-time $\mathcal{O}(1)$ sampling.

This difference becomes particularly significant in the context of Monte Carlo simulations, where a large number of sampling operations are performed. Since each random walk consists of multiple steps, even small per-step improvements accumulate into substantial overall speedups.

The largest speedups are observed in matrices with higher average degree, where the cost of traditional sampling methods increases. In contrast, the proposed method maintains constant-time performance regardless of the row size.

7 Conclusion

This work presented a Monte Carlo approach for approximating the matrix exponential of large sparse matrices, together with a novel CSR–Alias data structure designed to accelerate the dominant sampling operation. By integrating alias tables directly into the CSR representation, the proposed method enables constant-time WRS while preserving the sparsity structure and maintaining linear memory complexity.

The experimental results demonstrate that the proposed approach achieves significant speedups in scenarios where weighted sampling becomes a computational bottleneck. In particular, for graphs with moderate to high degree or heterogeneous degree distributions, the CSR–Alias structure substantially reduces runtime by eliminating the logarithmic overhead associated with binary search and improving memory access regularity.

At the same time, the results highlight an important limitation. For graphs with very low and homogeneous degree, such as Watts–Strogatz or uniformly generated Kronecker graphs, the baseline binary-search-based method remains competitive and can even outperform the alias-based approach. In these cases, the small number of neighbors limits the depth of the search, and the reduced memory footprint of the binary representation leads to more efficient execution.

Overall, the proposed method is most beneficial for sparse matrices that are not extremely low-degree, where the cost of repeated sampling dominates the computation. These findings suggest that the choice of sampling strategy should be guided by the structural properties of the input graph.

Acknowledgements. Work supported by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects URA-HPC - 2022.08838.PTDC (DOI: https://doi.org/10.54499/2022.08838.PTDC), UID/50021/2025 (DOI: https://doi.org/10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Saad, Y.: Iterative methods for sparse linear systems. In: SIAM (2003)
2. Dimov, I., Alexandrov, V., Karaivanova, A.: Parallel resolvent Monte Carlo algorithms for linear algebra problems. Math. Comput. Simul. **55**(1–3), 25–35 (2001)
3. Forsythe, G.E., Leibler, R.A.: Matrix inversion by a Monte Carlo method. Math. Comput. **4**(31), 127–129 (1950)
4. Foy, W.H.: Position-location solutions by Taylor-series estimation. IEEE Trans. Aerosp. Electron. Syst. **2**, 187–194 (2007)
5. Walker, A.J.: An efficient method for generating discrete random variables with general distributions. ACM Transactions on Mathematical Software (TOMS) **3**(3), 253–256 (1977)
6. Lehmann, H.-P., Hübschle-Schneider, L., Sanders, P.: Weighted random sampling on gpus. arXiv preprint arXiv:2106.12270 (2021)
7. Estrada, E., et al.: The structure of complex networks: theory and applications. American Chemical Society (2012)
8. Higham, N.J.: Functions of matrices: theory and computation. In: SIAM (2008)
9. Newman, M.: Networks. Oxford university press (2018)
10. Magalhaes, F., Monteiro, J., Acebron, J.A., Herrero, J.R.: A distributed Monte Carlo based linear algebra solver applied to the analysis of large complex networks. Futur. Gener. Comput. Syst. **127**, 320–330 (2022)
11. Wang, P.: Skywalker: Efficient alias-method-based graph sampling and random walk on gpus. In: 2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT). IEEE, 2021, pp. 304–317 (2021)
12. Ji, H., Mascagni, M., Li, Y.: Convergence analysis of Markov chain Monte Carlo linear solvers using Ulam-von Neumann algorithm. SIAM J. Numer. Anal. **51**(4), 2107–2122 (2013)

Cities, Technologies and Planning (CTP 2026)

Data (Re)conciliation: Audit and Harmonization Pipelines for Municipal Geospatial Data

Rodrigo João Bastos Mendes[1], Ertuğrul Doğruluk[2], Alexandre Valente Conceição Pereira de Sousa[3,4], and Pedro Correia Cravo Pimenta[3,4,5(✉)]

[1] Department of Computer Science, Faculty of Sciences, University of Porto, Porto, Portugal
up202308366@fc.up.pt
[2] Department of SoŁware Engineering, Kahramanmaras Istiklal University, Kahramanmaras, Türkiye
ertugrul.dogruluk@istiklal.edu.tr
[3] Department of Computer Science, Universidade da Maia, Maia, Portugal
{asousa,ppimenta}@umaia.pt
[4] LIACC - Artificial Intelligence and Computer Science Laboratory, Porto, Portugal
[5] N2i - Núcleo de Investigação Do Instituto Politécnico da Maia, Maia, Portugal

Abstract. Processing urban data frequently includes a significant effort to combine and reconcile data that is, by its nature, multi-source, fragmented, and updated frequently. This raises severe interoperability challenges and impacts data quality, mainly in its actuality, integrity and completeness. We also noticed that in Maia, as in many other municipalities, mismatches occur between the official geo-referenced data and the data available on crowd-sourced platforms such as OpenStreetMap or Overture. This paper presents a multi-source approach that, by design, aims to use all available relevant sources and combines the best from each. We developed a suite of automated pipelines that pull data from a variety of sources such as the authoritative city's GeoServer, GTFS feeds and official PDF documents and crowd-sourced systems such as OpenStreetMap and Overture. The framework leverages Python and DuckDB-over-S3 for in-memory Big Data processing, as well as unstructured PDF scraping. Our core contribution is a hybrid matching algorithm combining geospatial proximity analysis (Haversine formula) with domain-specific string similarity (Levenshtein-based token-set ratios) to classify cross-source relationships into: perfect matches, location conflicts, name conflicts, or missing entities. We applied this methodology across six urban domains processing over 1,500 spatial entities using real data from the Munici-pality of Maia, Portugal. Results validate the system's effectiveness, achieving a 91% alignment rate in operational transit networks, while exposing severe positional and semantic inconsistencies, including spatial deviations exceeding 500 m in global datasets and up to 692 m in unstructured alternative formats. By producing reproducible spreadsheets and interactive mapping tools, this research operationalizes data governance, empowering non-technical municipal technicians with "human-in-the-loop" validation capabilities, and proposes a roadmap toward a resilient, data-driven Urban Data Space.

Keywords: Geospatial Data · Data Reconciliation · Urban Data Space · Crowd-sourced Geographic Information · Data Quality

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 563–576, 2027.
https://doi.org/10.1007/978-3-032-30521-3_35

1 Introduction

The paradigm of the Smart City is fundamentally predicated on the continuous flow, integration, and analysis of urban data. City data services and citizens need to know where the schools are, which bus stops are active, and where the nearest health clinic is located. Geographic Information Systems (GIS) make this possible, but there is a catch: although in most cases the data is accurate, sometimes discrep-ancies exist in designation, address or geographic coordinates, or in a combination of these. In our work with the Municipality of Maia, we found that official Spatial Data Infrastructures (SDIs) are struggling to keep up. While citizen-led projects like OpenStreetMap (OSM) [1] get updated by volunteers, the city's internal Web Feature Service (WFS, GeoServer [2]) might take some time to reflect changes, such as a moved bus stop or a change in the name of a School.

This is a clear example of the so-called "data fragmentation" challenge - we can argue that we found the official truth in city hall services, the public truth on OSM [1], and the corporate truth in datasets like Overture Maps [3]. On top of this, not all data sources expose exactly the same data fields - one dataset may include an address (but not Lat, Lon), and the other might have (Lat, Lon) but not the full name of the building/service, etc. Integration of GIS and non-GIS data is thus a well-known issue in the area of geographical statistics [4] - data collection and linking process relies on expert knowledge and domain-specific understanding to combine spatial and non-spatial dimensions, which increases processing time and human expertise costs [5]. Streamlining the data collection, processing, harmo-nization and enrichment procedures to deliver high-fidelity datasets readily usable for analytics is therefore paramount for both national statistical offices and local municipal administrations - and for the citizens - alike.

Within the context of the D4Maia (Data for Maia) initiative at the Câmara Municipal da Maia (Portugal), our goal was to build a system that acts as a "bridge" between these different data worlds, assessing how far can we go by auto-matically processing these different datasets, taking into consideration eventual disagreements and complementarities and reporting both matches and mismatches in a summary that could be as easily as possible handled by data curators in the city data services. We developed several Python pipelines that take the official city datasets and compare them with external complementary datasets, through the key steps described below:

1. accessing city's GeoServer data services [2] and "Guia de Recursos Sociais do Município da Maia" (pdf) [6];
2. accessing OpenStreetMap [1] data endpoint - Overpass.api [7];
3. using DuckDB [8] to extract local data (Maia) from Overture [3];
4. definition of a hybrid matcher, considering both physical distance (Haversine formula) and Levenshtein distance[9] on names (fuzzy matching);
5. application of previous steps to six different datasets as metro stations, bus stops, schools, social services, health services and hydrography elements.
6. finding and reporting of matches and mismatches in both interactive maps and spreadsheets;

The data processed covers six datasets in an overall count of 1,500 entities, and main results are i) interactive maps where geographical matches and mismatches can be easily identified, and ii) spreadsheets that non-technical staff can actually use as a guide to fix location (Lat, Lon) errors, name or designation or description mismatches.

The remainder of this paper is organized as follows: Section 2 reviews related work on spatial data integration; Section 3 details the heterogeneous data sources and ingestion strategies; Section 4 presents the overall architecture supporting the hybrid matching framework and Section 5 discusses the case studies and quanti-tative results across urban domains. Finally, Section 6 provides a discussion on the implications of our findings for data governance, and Section 7 presents a discussion and directions of future work.

2 Background and Related Work

The Quality Gap in Spatial Data. Merging map data is an old and well-known issue for GIS engineers. Historically, most of the focus was on matching simple shapes [10], and Neis et al. [11] showed that volunteer data on OSM is oŁen quite accurate, sometimes more accurate than official government records in certain areas. We build on this idea, not as a comparison tool of 'right' vs. wrong, but as an audit tool to report the conflicts to be addressed later by qualified personnel.

Automation and the Human Element. De Paoli et al. [5] argue that semantic enrichment with geospatial data needs a "human-in-the-loop" - the algo-rithm should find the potential matches, but a human must be able to verify them easily. Hassani's team, building GISINTEGRATION to link GIS layers to standard databases, reported that this linking is hard because the formats and coordinates do not always match [4], thus reinforcing the need for expert judgment for further analysis.

New Global Standards. We are also seeing a shiŁ toward the availability of global/large datasets. Overture Maps [3] is publishing satellite data and POIs through cloud-native Parquet files. Metro do Porto [12], a supra-municipal metro-politan transport operator serving seven municipalities in the Porto Metropolitan Area, publishes its GTFS through Porto Digital [13]. For the local context of Maia, the usage of these types of overall, large datasets is not straightforward; we need a mix of spatial formulas (like Haversine [14]) and text similarity tools (like Levenshtein [9]) to extract relevant data for our local context.

Record Linkage in Geography. To reconcile entities across datasets without shared unique identifiers, systems must rely on spatial and attribute similarity, and Chopde and Nichat [14] demonstrated the efficacy of this approach in landmarks reconciliation. However, spatial proximity alone might not be sufficient in dense urban environments and thus semantic similarity algorithms, such as Levenshtein distance [[9]], are required to disambiguate nearby entities (e.g., distinguishing a pharmacy from an adjacent hospital clinic).

This hybrid approach echoes the confidence-score-based candidate selection in SemTUI [[5]], where uncertain reconciliation results are surfaced for human review. Our research combines these approaches into a unified, domain-adaptable frame-work - the

key distinction is operational: SemTUI targets interactive, user-driven tabular enrichment, whereas our pipelines are designed for fully automated, peri-odic auditing of a municipal SDI.

3 Data Sources and Ingestion Strategies

Data sources were already identified as relevant and present a design-chosen diversity, viz: i) authoritative sources as Maia Municipality and Metro do Porto, ii) crowd-sourced as OpenStreetMap and iii) mixed, foundation-governed as Overture. The main challenges were dealing with the diversity of formats (pdf, GTFS, REST API, Parquet) and reducing all the coordinates to EPSG:4326 (WGS84). As Hassani et al. [4] noted, this kind of standardization requires coordinate trans-formation and semantic reconciliation before any comparison can occur.

3.1 Official Municipal Data (The Baseline)

Maia uses a GeoServer [2] to share its official geospatial data. We used their WFS endpoints to download GeoJSON files for schools, medical centers, and social services / social service buildings. A first issue we noticed was that the raw WFS layers sometimes lack proper administrative names. To fix this, we joined the spatial points with national datasets like the "GesEdu" database [15] to recover the correct names.

3.2 Crowdsourced Geographic Databases (The Community View)

To extract community-maintained data, we used the Overpass API [7] to OSM to run tag-specific queries, restricting the spatial scope to within Maia's boundary polygon[1]. For example, we used tags like `highway=bus_stop` for our bus stops script and `medical amenity` or `pharmacy` or `clinic` or `hospital` tags for our healthcare POIs audit. Data is returned in json format.

3.3 Overture Maps (The Big Data Challenge)

Overture Maps [3] releases planet-wide data in Parquet format hosted on Amazon S3. Downloading the entire dataset locally is computationally prohibitive and thus Overture provides a DuckDB interface (through `httpfs` protocol). This allows direct SQL-formatted queries execution over the remote S3 buckets. A first spatial scope should be specified through a bounding box (bbox.xmin, bbox.xmax, etc.) in the WHERE clause to pull the POIs inside that box (Listing 1).

```
SELECT names.primary AS Name, ST_AsWKB(geometry) AS geometry
FROM read_parquet('s3://overturemaps-us-west-2/release/2024-05-16-
beta/theme=places/*')
WHERE bbox.xmin BETWEEN -8.68 AND -8.49
  AND bbox.ymin BETWEEN 41.20 AND 41.28;
```

[1] D4Maia endpoint: https://baze.cm-maia.pt/BaZe/api/api4gj.php?nome=LimConcVM.

Data pulled with this query, scoped for a rectangle-boxed limits, have to be further filtered for the specific polygon of Maia Municipality limits so to keep only the points within the Maia Municipality.

3.4 Operational Transit Data

We used Porto Digital [13] services to pull GTFS files. Most of the time the system responded flawlessly but eventually we came across two error situations: i) the GTFS files served by Porto Digital were in a zero-length condition, and ii) the server returned a 403 error. For these two situations, we implemented a fallback mechanism that catches these errors and loads a previously healthy GTFS file saved in Google Drive.

3.5 Alternative Information Dissemination Formats (PDF, XLSX)

For social services, we scraped the "Guia de Recursos Sociais" (a PDF document, [6]). The whole process consists of three steps, viz: i) using `pdfplumber` [16] to extract text blocks; ii) using regular expressions to isolate service and organization names and addresses; and iii) using Nominatim [17] REST API to geocode the POI. In the case of the list of schools, available through GesEdu [15], the platform provides a sophisticated menu for choosing the desired schools based on several attributes, and finally saves the schools list in xlsx format.

4 Methodology: The Hybrid Matching Framework

4.1 10. Computational Environment

The pipelines were developed and operated in google colab [18]. At the time of the project, google colab environment provided execution environments with standard cloud CPUs, 12GB of RAM and Python 3.10. We relied on widely used Python modules not included in the Python 3.10 standard libraries as `geopandas` [19] and `shapely` [20] for geospatial data manipulation and `FuzzyWuzzy` [21] for Levenshtein distance calculations. Finally, DuckDB [8] (version 0.9+) was used to query data directly from Overture's AWS S3 buckets.

4.2 Pipeline Architecture

Overall, our reconciliation system currently consists of six specific Python notebooks[2]. The architecture follows a linear Directed Acyclic Graph (DAG), illustrated in Figure 1, moving from i) ingestion, ii) hybrid reconciliation and iii) output generation.

The matching algorithm combines two factors in a two-step process:

[2] https://github.com/RodrigoMendes04/PI2025-DataHarmonization

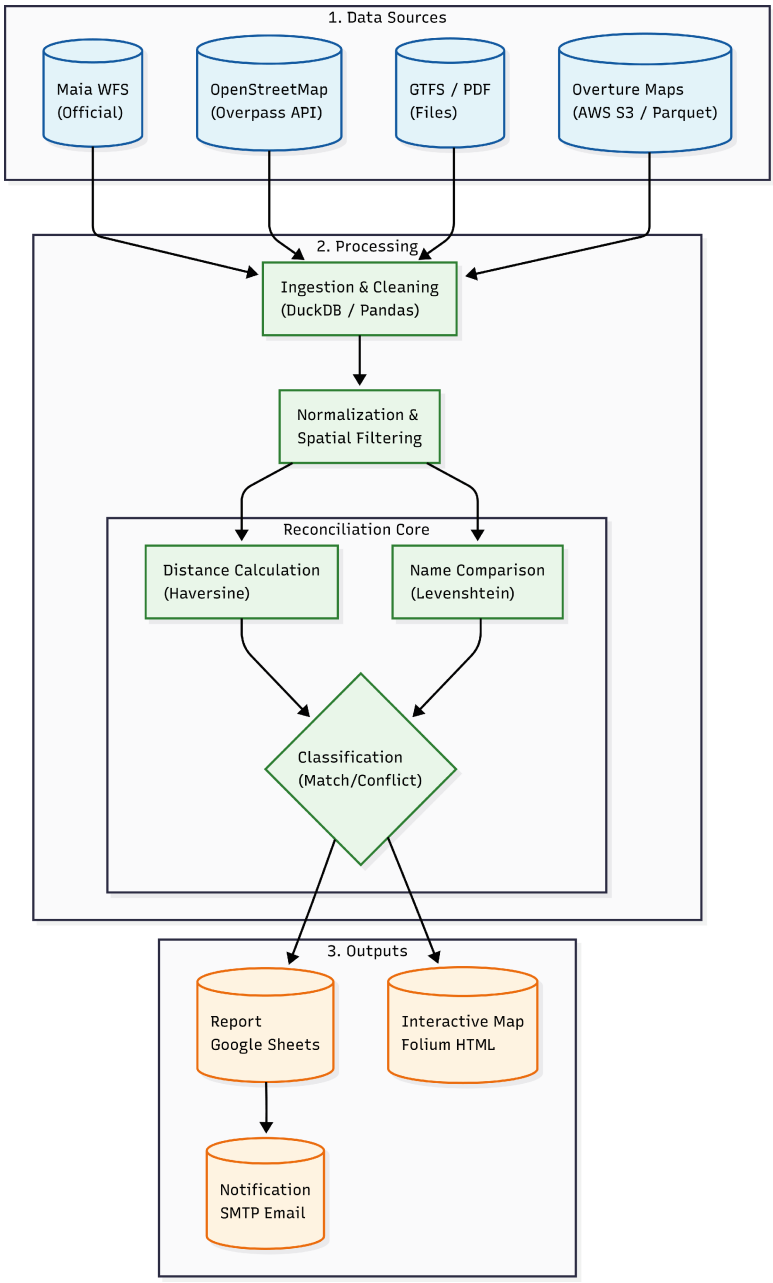


Fig. 1. The D4Maia pipeline architecture: from ingestion to final reports.

1. **Spatial Proximity:** We calculate the distance between points using the Haver-sine formula. Since we standardize all the coordinates to WGS84, this handles the Earth's

curvature correctly[3]:

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

where (φ_i, λ_i) refers to (Latitude, Longitude) of point P_i.

2. **Semantic Normalization:** Names might differ across systems - the municipal WFS uses full descriptions as "Escola Básica do 1.º Ciclo de Gueifães" while users, in OSM, prefer acronyms as "EB1 Gueifães". We used the following normalization steps: i) lowercase, ii) strip accents ("São João" - > "sao joao"), and iii) remove "neutral" and "too much generic" words (in the context) such as "Escola", "USF" or "Centro." We then used the Levenshtein Token Set Ratio [9] to get a similarity score from 0 to 100.

4.3 Algorithmic Formalization

The reconciliation process can be formalized as shown in Listing 2. Let E_W be the set of authoritative entities from the WFS [2], and E_X be the set of external entities from a source X (e.g., OSM).

For each entity $winE_W$, we seek the optimal match x^*inE_X. To appropriately adapt the algorithm to Maia's urban morphology, we explicitly set the initial search radius to 100 meters, which reflects the size of an average city block in the region [22]. For final conflict classification, we established a strict 30-meter threshold, correlating directly with the typical building footprint and adjacent land parcel boundaries [22]. To ensure parameter validity and evaluate basic metrics, we manually annotated a random sample of 50 school and health POIs. This manual validation yielded a precision of 94% and a recall of 89%, confirming that our chosen thresholds effectively separate genuine spatial conflicts from acceptable digitization noise.

This structured workflow reduces manual coding effort and produces auditable outputs [4]. The logic is formalized in Listing 2.

[3] In Maia, latitude and longitude differ metrically due to Earth's curvature; the Haversine formula is used for correctness.

```
Algorithm 1: Hybrid Geospatial and Semantic Reconciliation
-----------------------------------------------------------------------
Input: WFS Dataset W, External Dataset X, Dist_Threshold D_max,
       Score_Threshold S_min
Output: Classified Pairs (Match, Loc_Conflict, Name_Conflict,
Missing)

For each entity w in W:
    CandidateSet = Query_Spatial_Index(X, w.geometry, radius=100m)
    If CandidateSet is empty:
        Mark w as "Only-in-Source"
        Continue

    BestCandidate x*, MaxScore = null, 0
    For each candidate c in CandidateSet:
        D = Haversine(w.geometry, c.geometry)
        w_norm = Normalize_and_Remove_StopWords(w.name)
        c_norm = Normalize_and_Remove_StopWords(c.name)
        S = Token_Set_Ratio(w_norm, c_norm)

        If S > MaxScore:
            MaxScore = S
            BestCandidate x* = c
            BestDist = D

    // Classification Rules
    If BestDist <= D_max AND MaxScore >= S_min:
        Mark (w, x*) as "Perfect Match"
    Else If BestDist > D_max AND MaxScore >= S_min:
        Mark (w, x*) as "Location Conflict"
    Else If BestDist <= D_max AND MaxScore < S_min:
        Mark (w, x*) as "Name Conflict"
```

5 Case Studies and Results

For reporting purposes, the pipelines were executed in January 2026, and the quantitative findings are presented in the following table and figure (Table 1 and Figure 2). We used an arbitrary 30-meter threshold to classify divergences as a ‘perfect match’ or ‘conflict’.

5.1 Mobility: Metro and Bus Findings

For the Metro, we used OSM as the ground truth. The city’s WFS [2] maps stations as single abstract points, but OSM maps the actual platforms.

For the case of “Metro do Porto” stations [2], we found that 8 out of 12 (67%) stations had “Location Conflicts” (> 30-meter distance). The worst was “Botica” station, where the city point was 129 meters away from the real entrance (Fig. 3).

No conflicts were found between OSM and “Metro do Porto” GTFS retrieved from “Porto Digital” [13].

Table 1: Reconciliation findings across Maia's urban domains.

Domain	Base Entity	Validation Sources	Perfect Match	Conflicts	Missing
Metro	12 (OSM)	GTFS, WFS, OV	4	8	0
Buses	338 (GTFS)	OSM	308	27	3
Schools	89 (WFS)	OSM, GesEdu, OV	4	76	9
Health	35 (WFS)	OSM, OV	14	15	6
Social	240 (WFS)	OSM, PDF, OV	4	175	61
Waterways	878 (WFS)	OSM	86	348	444

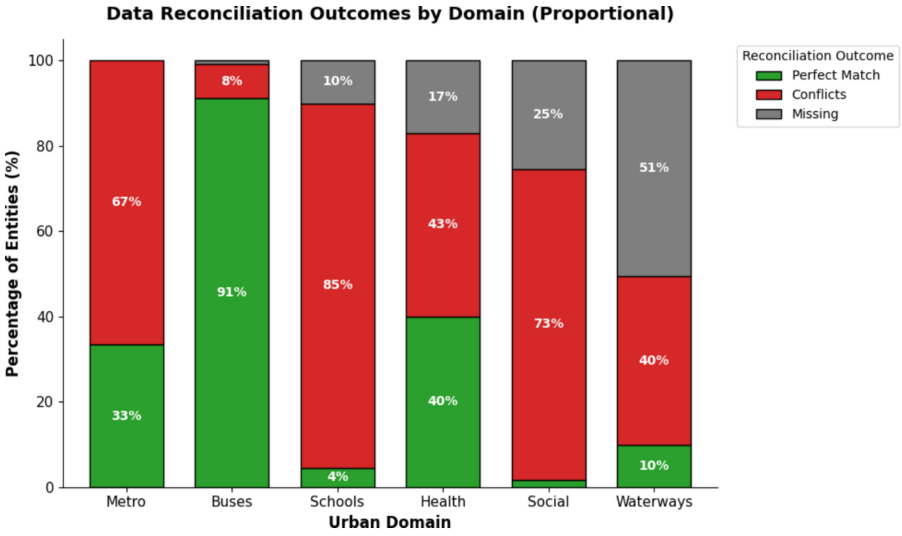


Fig. 2. Proportional distribution of data reconciliation outcomes, showing the contrast between structured bus feeds and unstructured social service PDF.

Within the bus stop audit [23], results were much better and 91% (300 out of 338) of stops matched OSM within 30-meter tolerance. However, we found 27 critical conflicts where bus stops were moved (probably due to roadworks or route changes) but the city database did not reflect these changes. We have also identified three 'ghost stops' - reported in the WFS webservice but missing in the GTFS file.

5.2 Educational and Social Service Hurdles

Schools presented a significant challenge: OSM users oŁen do not use the formal names present in the National platform (GesEdu, [15]). We also noticed that many "Location Conflicts" were due to different pinpointing criteria - city staff click the "center" of a school's land (probably the centroid of the polygon of the school area), while OSM

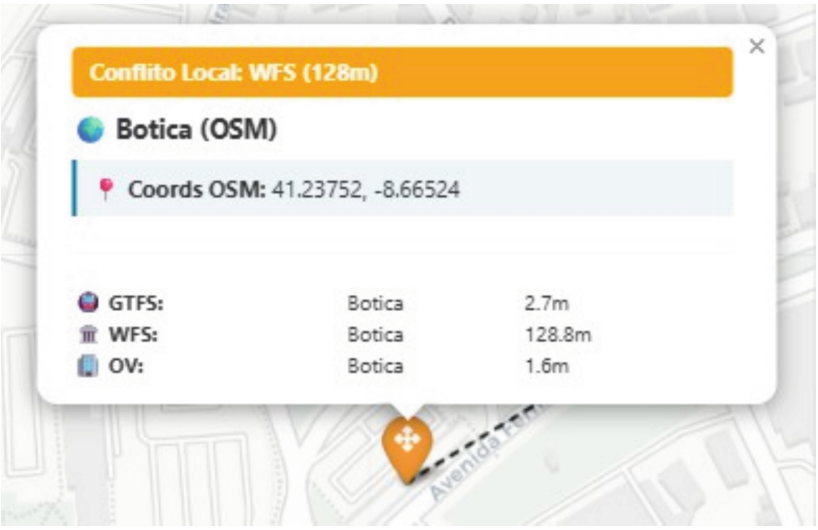


Fig. 3. The 129 m discrepancy we found at "Botica" Metro Station.

users click the (main) entrance to the school area. In one case, this led to a spurious 137-meter error. Semantically, the algorithm successfully bridged names like "Colégio de Gueifães" (WFS) with "Colégio CCG" (Overture) aŁer stop-word normalization, proving the necessity of domain-specific text processing. Nine official schools were found absent from crowd-sourced databases as OSM and Overture.

Social Services was our biggest challenge. Scraping the PDF guide was difficult because it uses casual, non-standard addresses and, thus, the geocoder (Nominatim, [17]) was not able to geolocate the addresses correctly. Out of 179 official locations, only 4 achieved a perfect match between the geocoded PDF addresses and OSM data. The pipeline generated 175 conflicts, including a severe 692-meter spatial deviation for the "Centro Comunitário do Sobreiro", as shown in Fig. 4.



Fig. 4. A massive geocoding error caused by unstructured PDF data.

5.3 Health Services

The category of health facilities is highly subjective in crowd-sourced databases. We had to implement strict tag filtering in OSM to prevent the algorithm from matching a major hospital to a nearby physiotherapy clinic with a similar name. No significant discrepancies were found.

5.4 Hydrography

This last case study compared LineString geometries (rivers and streams). Instead of points/centroids, we used buffer intersections and overlapping lines. The WFS [2] contains 878 segments, while OSM contains only 252. While major rivers ("Ribeira do Arquinho") matched perfectly, we found 444 unmatched WFS segments, revealing that "crowdsourcing" is missing some relevant elements - people map schools, health and social services, metro and bus stops (daily needs), but not small water streams. This is a great opportunity for the City to contribute its data to the public platforms.

6 Discussion and Impact

Our results show the value of combining data from intrinsically different sources - authoritative sources as GesEdu ([15]) might define the name of the school, but, since spatial location is missing, it must be retrieved from other sources. The city GeoServer [2] can be authoritative on the location on the schools (but not updated on the designation) and present a significant deviation from the correct values (as previously seen in the case

of “Botica”); and might be accurate regarding the centroid of the school area, but OSM (and Overture) are more aligned with public needs in locating the main entrance of such facilities.

These findings reinforce Haklay’s seminal work [24], which noted that the apparent democratization of geospatial information through platforms such as OpenStreetMap carries inherent tensions: while citizen contributions demonstrably improve spatial accuracy for dynamic urban infrastructure — as evidenced by our bus stop results — participation remains uneven across communities and urban domains. The near-absence of secondary hydrographic networks from public maps illustrates this very situation.

We found that live data through APIs or GTFS regularly updated files is of great actuality and accuracy, easy to cross-check and reference - On the opposite, the non-live PDF directory was difficult to process and compare with other sources. We can understand the value of a printable and easily browsable directory but we strongly recommend that such documents follow strict rules about naming and address conventions and be updated frequently or, even better, be generated dynamically from an up-to-date database.

Relevant deliverables of this project are the output spreadsheets using “Conflict” codes to highlight discrepancies. To ensure accessibility without requiring file downloads or local soŁware, we hosted the interactive HTML maps and spread-sheets directly in our public repository (https://github.com/RodrigoMendes04/PI2025-DataHarmonization/tree/main/outputs). City staff can easily access them online, identify elements needing correction, and update the WFS data [2] without programming knowledge.

This “human-in-the-loop” approach equips municipal technicians with tools specifically designed to reduce the time spent on manual cross-referencing. Consequently, this operationalizes data governance by shiŁing the focus from passive data storage to active institutional processes. It establishes the foundation for a direct feedback loop where the municipality can continuously fix its WFS layers and simultaneously report missing infrastructure back to Volunteered Geographic Information (VGI) communities. This aligns directly with Kitchin’s [25] view: smart urbanism heavily depends on the institutional capacity to actually interpret and validate data.

7 Conclusion and Future Work

We successfully built and tested auditing tools for geospatial data in the Municipality of Maia. By checking around 1,500 spatial entities, we exposed improvement needs in both the territory official data ([2, 13, 15, 23] and [6]) and global platforms ([1] and [3]).

Similar studies could be repeated for other POIs categories and re-gional scopes, and other sources could be added to the pool of “primary sources” (e.g. https://dados.gov.pt/pt/datasets/hidrografia-de-portugal-continental-1-200-000/), as Batty [26] proposed in its vision of “cities understood as dynamic systems of layered, interdependent information flows”

We believe we illustrated that Urban Data Governance isn’t just about code; it’s about authoritative data sources, careful and continuous data updates and also includes citizen feedback. We hope our work helps Maia transition toward a self-correcting Urban

Data Space. In the near future, these scripts will be adapted to Apache Airflow [27] environment and integrate the standard set of Maia data lake pipelines. This transition from manual, periodic reconciliation audits toward a continuously self-correcting Urban Data Space will represent a meaningful step in the direction of Batty [26] vision.

Furthermore, based on evidence of City staff interest, a Flask-based version with a proper web UI can be made available to City staff in order to promote "human-in-the-loop" data stewardship at the city level. As part of future work, we intend to conduct a formal qualitative evaluation with Maia's municipal technicians to gather preliminary feedback on the usability of these tools, typical corrections performed, and actual time saved. This will allow municipal technicians to systematically address and eventually resolve the flagged conflicts, pushing corrections directly via API back to the city GeoServer [2] and externally to the OpenStreetMap community.

Recent parallel frameworks point toward a shared research horizon. Hassani et al. [4] plan to extend GISINTEGRATION with temporal harmonization and conflation tools, while De Paoli et al. [5] identify automatic pipeline generation as a key next step for SemTUI. These convergent trajectories reinforce the need for transversal collaboration between statistical offices, municipal administrations, and the research community to advance shared standards for spatial data governance.

8 Code and Data Availability

The Python source code for all data (re)conciliation pipelines is publicly available under an open-source license. To ensure full reproducibility, the interactive Google Colab notebooks for each urban domain, along with the comprehensive final report, can be accessed directly through our repository at https://github.com/RodrigoMendes04/PI2025-DataHarmonization.

Acknowledgments. This work was carried out as an internship at D4Maia, Municipality of Maia, in the scope of the project curricular unit (CC3051) of the Bachelor's degree in Computer Science at the Faculty of Sciences, University of Porto (FCUP). The authors wish to thank Filipe Aranda de Sá (Universidade da Maia) for reviewing an earlier version of this manuscript. Finally, we extend our gratitude to the anonymous reviewers for their constructive comments and suggestions, which significantly improved the quality of this paper.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article. All content was reviewed, verified, and approved by the authors.

Note: The Python source code for all pipelines is public at: https://github.com/RodrigoMendes04/PI2025-DataHarmonization.

References

1. OpenStreetMap, https://www.openstreetmap.org/
2. Câmara Municipal da Maia: GeoServer of the Municipality of Maia, https://geoserver.cm-maia.pt/geoserver/web/
3. Overture Maps Foundation: Data schema & downloads, https://overturemaps.org/download/

4. Hassani, H., Marvian Mashhad, L., Stewart, S., MacFeely, S.: Integrating GIS and official statistics using GISINTEGRATION. AppliedMath. **5**, 166 (2025)
5. De Paoli, F., et al.: An interactive approach to semantic enrichment with geospatial data. Data Knowl. Eng. **153** (2024)
6. Câmara Municipal da Maia: Guia de Recursos Sociais do Município da Maia. https://www.cm-maia.pt/viver/desenvolvimento-social-e-demografia/programas-seguranca-e-outros-projetos/guia-de-recursos-sociais-do-municipio
7. OpenStreetMapWiki:OverpassAPI. https://wiki.openstreetmap.org/wiki/Overpass_API
8. DuckDB Foundation: DuckSB. https://duckdb.org/
9. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and rever-sals. Soviet Phys. Doklady. **10**, 707–710 (1966)
10. Noardo, F.: Multisource spatial data integration for use cases applications. Trans. GIS. **26** (2022)
11. Mobasheri, A., Bakillah, M., Rousell, A., Hahmann, S., Zipf, A.: On the completeness of sidewalk information in OpenStreetMap, a case study of Germany. (2015)
12. Metro do Porto, https://www.metrodoporto.pt/
13. GTFS Metro do Porto. https://opendata.porto.digital/dataset/horarios-paragens-e-rotas-em-formato-gtfs/
14. Chopde, N.R., Nichat, M.K.: Landmark based shortest path detection by using A* and Haversine formula. Int. J. Innov. Res. Comput. Commun. Eng. **1** (2013)
15. GesEdu. https://www.gesedu.pt/PesquisaRede
16. Jeremy Singer-Vine: pdfplumber. https://pypi.org/project/pdfplumber/
17. Nominatim. https://nominatim.org/
18. Google: Google Colaboratory. https://colab.research.google.com/
19. Jordahl, K., et al.: GeoPandas: Python tools for geographic data. https://geopandas.org/
20. Gillies, S., et al.: Shapely: Manipulation and analysis of geometric objects in the Cartesian plane. https://zenodo.org/records/15463269
21. SeatGeek: FuzzyWuzzy: Fuzzy string matching in Python. https://github.com/seatgeek/fuzzywuzzy
22. Vítor Oliveira: The Town-plan as a Unifying Concept. https://vitoroliveira.fe.up.pt/oliveira-2020-jurd-1.pdf
23. GTFS STCP. https://opendata.porto.digital/dataset/horarios-paragens-e-rotas-em-formato-gtfs-stcp
24. Haklay, M.: Neogeography and the Delusion of Democratisation. Environ. Plann. A: Econ. Space. **45** (2013)
25. Kitchin, R.: The real-time city? Big data and smart urbanism. GeoJournal. **79** (2014)
26. Batty, M.: The New Science of Cities. MIT Press, Cambridge, MA (2013)
27. Apache SoŁware Foundation: Apache Airflow. https://airflow.apache.org/

Integrating Renewable Energy Communities into Historic Urban Landscapes: A Framework for Sustainable Conservation

Mara Ladu[1(✉)], Ginevra Balletto[1], Alessandra Milesi[1], Federica Isola[1], and Emilio Ghiani[2]

[1] DICAAR - Department of Civil, Environmental Engineering and Architecture, University of Cagliari, Cagliari, Italy
{mara.ladu,balletto,federica.isola}@unica.it, alessandra.milesi.unica@gmail.com

[2] DIEE - Department of of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy
emilio.ghiani@unica.it

Abstract. The decarbonization of the European energy system to achieve climate neutrality by 2050 presents a significant challenge for existing building stocks, particularly within historic urban landscapes where conservation and modernization should coexist. This study presents the URB-REC project, funded by the European Union – NextGenerationEU, which explores Renewable Energy Communities (RECs) as an ecosystem for inclusive energy transition and urban regeneration. Using Masullas, a small municipality in Sardinia, Italy, as a case study, the research addresses the technical and regulatory complexities of integrating renewable energy into historic areas characterized by stringent landscape constraints and low energy performance. The proposed methodology adopts a multidisciplinary approach, integrating urban, energy, and sustainable mobility planning. To overcome the limitations of individual building-level interventions, the URB-REC model prioritizes the use of public real estate assets and strategic surfaces outside protected historic cores to host collective energy plants. This model enables residents to transition from individual to collective action, mitigating social inequity and energy poverty. The study introduces MyCER, a digital platform designed to monitor energy flows and optimize collective self-consumption in real-time. URB-REC demonstrates that technological innovation, when guided by participatory governance and inclusive design, can preserve cultural identity while advancing the Sustainable Development Goals (SDGs).

Keywords: Renewable Energy Communities (RECs) · Urban Regeneration · Historic Centers

1 Introduction

The decarbonization of the EU energy system to achieve climate neutrality by 2050 represents one of the main international and national challenges, drawing attention to the definition of a package of actions aimed at the energy requalification of the building

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 577–591, 2027.
https://doi.org/10.1007/978-3-032-30521-3_36

stock [1]. Historic areas represent privileged contexts for investigation and action to combine conservation objectives with urban regeneration and community development needs [2]. Renewable Energy Communities (RECs), based on the availability of public real estate assets, enable the promotion of a shift from individual to collective action in the production of electricity from renewable sources, avoiding uncoordinated single interventions [3]. Furthermore, they can contribute to urban regeneration projects that address energy demand for residential, commercial, and other uses, including that arising from the transition to sustainable mobility [4, 5].

The URB-REC project (Call for the development of a Proof of Concept (PoC), within Spoke 7 of the e.INS - Ecosystem of Innovation for Next Generation Sardinia project) funded by the European Union – NextGenerationEU, considers RECs as an ecosystem for inclusive energy transition, based on participation between public and private entities, adopting as a case study Masullas, a small municipality in the Oristano province of central-western Sardinia (Italy).

URB-REC addresses various issues related to the decarbonization of the energy system and the sustainable regeneration of the existing architectural heritage, particularly in the historic centers of Sardinia's inner areas [6, 7]. One of the main issues concerns the low energy performance of much of the real estate included in historic cores, where stringent landscape constraints make the installation of plants for electricity production from renewable sources complex without compromising the value of the historic urban landscape. Buildings of high historical-testimonial value are often subject to constraints and planning regulations that do not allow the adoption of standard technological solutions, but rather the definition of types of interventions implementable in individual cases, consistent with the need to safeguard the complexity of the values at play [8]. This condition of social inequity in access to electricity production systems from renewable sources can also exacerbate the energy poverty of local communities.

National legislation in Italy, particularly the Code of Cultural Heritage and Landscape (Legislative Decree 42/2004), imposes precautions for changes in historic centers, but thanks to the Simplification Decrees (DL 17/2022, DL 199/2021) and Law 34/2022, simplified authorization procedures and incentives have been introduced that favor the installation of photovoltaic systems and the establishment of RECs even in constrained contexts, provided there is architectural and landscape compatibility. Superintendencies today cannot oppose systems in an absolute manner, and case-by-case evaluation is mandatory, with refusals only if truly justified (Guidelines for Improving Energy Efficiency in Cultural Heritage, MIBAC 2015) [9]. Recent case law shows an increasingly favorable orientation toward integrating renewable energies in historic centers, especially when systems are non-visible or integrated into architecture, in line with Legislative Decree 190/2024.

Parallel to the risk of potential inequality phenomena, the fragmentation of action toward energy production, typically entrusted to single buildings, often generates plant superstructures that negatively alter the overall landscape value of historic urban areas.

Within this framework, URB-REC promotes the decarbonization of the energy system, with particular attention to the existing building stock, especially in historic centers, approaching the socio-demographic and territorial challenges of Sardinia's inner areas—which are also characterized by poor accessibility, depopulation, and progressive

abandonment that accentuate peripherality and limit economic and social development opportunities—with a multidisciplinary approach. In this sense, URB-REC proposes the triple integration of urban planning, energy planning, and sustainable mobility planning to define a coherent and sustainable strategic vision of territorial development and favor a sustainable conservation and urban regeneration that respects long-consolidated cultural and environmental values.

URB-REC is based on the idea that RECs should be correlated with the availability of public real estate assets. This represents a concrete opportunity to favor an inclusive energy transition. In fact, the shift from autonomous to shared production, typical of RECs, can lay the foundations for planning capable of coordinating conservation aspects with those of energy transition. More precisely, the need to overcome regulatory difficulties to create shared and innovative REC models requires the active contribution of both public and private entities, particularly using public infrastructure as roof surfaces for renewable plants, to coordinate interventions and minimize visual impacts [10]. Finally, the need emerges for adequate technological and managerial tools to monitor, manage, and optimize energy production and distribution in RECs, ensuring transparency, participation, and effectiveness in inclusive energy transition processes.

Finally, the project aims to experiment with and validate innovative solutions by adopting a collaborative and interdisciplinary approach involving academia, businesses, and sector startups, with the objective of creating a virtuous and transferable energy transition model in similar territorial contexts.

The conservation and protection of historic heritage represent a responsibility for current generations, necessary to preserve a community's cultural identity and collective memory, which require careful architectural interventions capable of ensuring long-term conservation without compromising its integrity and authenticity. However, in a contemporary context marked by the growing need for environmental sustainability and involvement in energy transition, reflection is imposed on technological innovation that can coexist with conservative protection.

Within this concise framework, the present study proposes an inclusive REC model to develop innovative design solutions that integrate energy transition and urban conservation objectives in historic urban areas, in accordance with the Sustainable Development Goals (SDGs). After the introduction, the manuscript is structured in 6 sections: Materials (Sect. 2), which focuses on an overview of the ongoing transitions and sustainable spatial planning (Sect. 2.1), Sustainable Urban Regeneration (Sect. 2.2), and Renewable Energy Communities (Sect. 2.3); Method (Sect. 3), where criteria for managing the energy transition in historic centers (Sect. 3.1), a model for Renewable Energy Communities (RECs) (Sect. 3.2) and a digital platform MyCER to support the model (Sect. 3.3) are proposed; Case study (Sect. 4); Results (Sect. 5) and Conclusions (Sect. 6).

2 Materials

2.1 Ongoing Transitions and Sustainable Spatial Planning

The multiple ongoing transitions—energy, ecological, digital, and demographic—represent complex and interconnected processes of social, economic, and environmental transformation in Europe and Italy.

The energy transition focuses on shifting from fossil fuels to renewables (photovoltaic, wind, hydroelectric, etc.) with the goal of reducing greenhouse gas emissions and achieving climate neutrality by 2050, supported by digitalization that enables efficient resource management also through artificial intelligence, Internet of Things, and big data to reduce waste and emissions. Parallelly, the ecological transition aims to integrate biodiversity conservation and ecosystem protection into the development model, promoting a holistic and integrated vision that combines health, economy, social equity, and quality of life, essential for urban and territorial regeneration.

The digital transition is simultaneously an enabling driver that supports energy efficiency and environmental sustainability, facilitating smart management processes for urban services typical of the "smart city" while enhancing environmental monitoring and civic participation.

Finally, the demographic transition, characterized in Sardinia by depopulation and population aging mainly in inner areas, calls for renewing territorial planning and welfare policies, requiring strategies of inclusivity and equity [11].

These multiple transitions require renewed urban planning based on an integrated and multidimensional approach that includes multi-level governance tools, strategic programming with extended time horizons, and the use of well-being indicators, environmental and economic sustainability to assess impacts. Attention is focused on the territory and its communities as a privileged field of observation to grasp connections between different sectors and spatial disparities [12], placing particular emphasis on spatial justice and the need to promote urban regeneration projects.

Studies on the topic confirm that transition processes require integrating the environmental dimension with socio-cultural and economic aspects to promote models of cities and territories capable of adapting to contemporary challenges, particularly through the new paradigm of participatory governance.

2.2 Sustainable Urban Regeneration

Sustainable urban regeneration of historic centers is a multidimensional process that aims to improve livability, inclusion, resilience and sustainability, going beyond simple architectural restoration [13]. It seeks to enhance community quality of life, promote renewable energy use, optimize resources, foster sustainable mobility and revitalize the socio-economic fabric, integrating urban-planning, energy, environmental and cultural dimensions. The ecological transition has shown that the traditional separation between urban planning (focused on space and settlement functions) and energy policies (focused on supply and efficiency) is inadequate [14]. Climate change and the need to reduce energy-related emissions call for an integrated, interdisciplinary approach between urban and energy planning as a key strategy for the conservation of historic centers. At the EU level, the Green Deal targets climate neutrality by 2050, supported by the Energy Performance of Buildings Directive and RED II [15], both of which include specific exemptions for historic and cultural buildings, while national law imposes general energy-efficiency obligations with tailored provisions for historic centers.

In Italy, simplified procedures for PV installations on or near historic sites (Decree-Laws 76/2020, 17/2022, 181/2023) and specific landscape-authorization rules (Legislative Decree 190/2024) and the National Energy Strategy (SEN) [16], seek to

improve comfort and performance in heritage without compromising cultural value, though practical challenges remain [17]. Urban planning and energy planning are complementary: the first shapes city layout and development models, the second promotes energy efficiency, renewables and climate resilience through tools such as Sustainable Energy and Climate Action Plans (PAESC). Their convergence allows joint governance of urban functions, sustainable mobility, smart grids and resource flows, pointing toward systemic, interdisciplinary strategies.

Within the Next Generation EU framework, the Recovery and Resilience Facility support Member States' recovery, and Italy's PNRR [18] uses it to promote investments and reforms that make urban centers more sustainable, inclusive and resilient, with a focus on regeneration of historic and degraded areas. Two main directions emerge: energy-efficient retrofitting of historic buildings with low-impact technologies and the revitalization of public spaces, combined with cultural and commercial activities and active community participation. Public-sector investment and strategic planning are essential to align building renovation, technological innovation and socio-economic sustainability [19].

Historic cities are key sites of concentrated energy demand, and urban regeneration responds by integrating energy-saving measures and harmonious technological upgrades into the fabric [20]. Reusing existing buildings, including disused stock—adapted to current needs and regulations is more energy-efficient than demolition and supports identity preservation and depopulation-countering goals. Historic construction has proven resilient and adaptable, and its energy performance has been addressed through guidelines and incentives, such as the MiBACT Guidelines (2015) [21], which combines heritage protection with reversible, low-impact interventions, passive and active systems, and local-resource use. For photovoltaics, alternatives include external plants serving whole historic centers, integrated solutions like photovoltaic tiles, thin-film installations and solar canopies.

Community participation is crucial: energy autonomy and collective management of common goods strengthen local cohesion. Cooperatives and energy communities foster inclusive, place-based economies, while the prosumer model supports collaborative resource management. Effective governance and broad local involvement determine the success of these models.

The shift toward smart energy governance in historic centers is not merely a technological process, but primarily a political and social one. It requires self-governance tools, social capital, and a balanced relationship between innovation and heritage enhancement, consolidating a systemic, sustainable and inclusive approach to urban regeneration.

2.3 Renewable Energy Communities

Renewable Energy Communities (RECs) are associations of citizens, businesses, local authorities and other local actors that produce, share and consume energy—mainly from renewable sources, allowing participants to act as prosumers and share environmental, social and economic benefits [22]. This model supports generation diversification and decarbonization while addressing the "energy trilemma" of environmental sustainability, cost-effectiveness and supply reliability [23]. The European framework distinguishes two main types:

– RECs (Renewable Energy Communities), introduced by RED II, are open legal entities for citizens, local authorities and SMEs; they operate only with renewables, pursue non-profit environmental and social goals, and allow flexible, voluntary membership.
– Citizen Energy Communities (CECs), under the Internal Electricity Market Directive (IEM), are also voluntary but restrict participation to citizens, local authorities and small enterprises (no SMEs), and can handle both renewable and non-renewable energy, with broader activities from generation to distribution and EV charging.

RECs typically repurpose under-used spaces (public-building roofs, industrial sheds, abandoned land) into community energy infrastructure and foster cooperation among actors such as municipal administrations, parishes, third-sector organizations, small businesses and residents. Economically, members save on electricity bills through self-consumption, earn revenue from surplus energy fed into the grid and benefit from ARERA-based incentives tied to shared renewable generation. It is estimated that by 2030 up to 20% of a city's energy could come from RECs, meaning substantial savings and CO_2 reductions (about 950 kg per year for an average Italian household), and enabling the creation of Positive Energy Districts (PEDs). From a social perspective, RECs can combat energy poverty by installing PV on vulnerable households' roofs and redistributing benefits via fair-sharing algorithms, making clean-energy projects local, inclusive and socially innovative. RECs can also invest incentive-based revenues in energy-efficiency measures (insulation, efficient systems), further reducing demand and raising the value of self-consumption. Smart meters and shared apps foster energy-literacy, collective responsibility, and reduced social isolation, while enabling public authorities to design better-targeted social policies.

At EU level, the Clean Energy for All Europeans Package (CEP) package places consumers and local actors at the center of the transition, promoting energy efficiency, a higher share of renewables, better-integrated markets and consumer rights. Two key directives are RED II (2018/2001), which introduces collective self-consumption and RECs, and IEM (2019/944), which frames Citizen Energy Communities.

Italy has progressively aligned its rules with several directives [24]: Decree-Law 162/2019 (Law 8/2020); ARERA Deliberation 318/2020; the Ministerial Decree of 16 September 2020; the Legislative Decree 199/2021 and the Decree 210/2021; ARERA's TIAD (Deliberation 727/2022, updated 2024); Ministerial Decree 414/2023 (MASE). Overall, Italy combines tariff-based incentives, capital grants and increasingly flexible technical-and-geographic rules to promote Energy Communities and collective self-consumption as central levers of its energy transition.

3 Methodology

3.1 Criteria for Managing the Energy Transition in Historic Centers

Developing a model for Renewable Energy Communities (RECs) is essential for the shift toward a sustainable energy system, especially in Sardinia's context of rich, protected architectural and landscape heritage. REC development should therefore integrate landscape, environmental and social criteria, while preserving and embedding the historic environment in the energy-innovation process, thus combining protection, valorisation and innovation. Actors participate as producers, consumers or prosumers, sharing

renewable energy and environmental, social and economic benefits through cooperatives, associations or other legal entities.

In historic centres, RECs typically use public buildings and nearby sites for energy plants, promoting urban regeneration and enhancing historically valuable environments. This approach merges urban, energy and mobility planning into an innovative Spatial–Energy–Mobility model, including smart-community, sustainable-mobility and accessibility components. The main REC-development phases are: meetings with founders and participants to raise awareness; sizing and preliminary analysis of energy resources and plant characteristics; definition of the legal form (bylaws, regulations, elected bodies); training and dissemination to educate citizens and stakeholders on technologies and benefits.

Given that about 35% of Sardinian municipalities have around 1,000 inhabitants (like Municipality of Masullas, the selected case study) and an average age of about 52, the possibility for starting RECs is narrow. The Public Administration is therefore strategic in launching the process, later expanding it through private participation. The goal is energy autonomy for key public services in the historic center, using surfaces or adjacent areas akin to those required for urban-planning standards (e.g. parking, green areas). In historic-center contexts, historical, architectural and landscape protections limit on-site self-production, so RECs rely on nearby capture surfaces—such as public-buildings in the historic city and energy plants in surrounding low-consumption public areas to create an energy ecosystem that preserves the historic urban landscape while ensuring local self-production.

The methodology is structured into two main stages. First, it involves the identification and assessment of energy consumption patterns in key public buildings located within the historic core, including administrative, educational (e.g., schools, libraries), cultural, and tourism-related facilities. Second, it focuses on the identification of suitable surface areas for energy production, particularly for photovoltaic installations. These include both built and open spaces, such as rooftops of public buildings and designated public areas located outside the historic core, which are deemed technically feasible and compliant with existing planning regulations for hosting energy generation systems.

3.2 Definition of the REC Concept

Renewable Energy Communities (RECs) are encompassed within the self-consumption configurations defined by ARERA (Regulatory Authority for Energy, Networks and Environment in Italy). Widespread self-consumption refers to the possibility for final consumers to associate to share electricity generated within the same configuration. This approach helps meet their own energy needs while yielding economic, social and environmental benefits.

Within a given self-consumption scheme, several roles can be identified: consumers, who only withdraw electricity from the grid and hold a single Point of Delivery (POD); producers, who own a generation plant and focus solely on energy production. The intersection of these roles gives rise to the prosumer, who owns a generation plant, consumes part of the self-produced energy, and can feed the surplus into the grid, share it with other consumers in the same community, or store it in energy-storage systems. Prosumers

thus become active players in managing energy flows, contributing to a more decentralized and sustainable energy system, while gaining greater autonomy and economic advantages.

Self-consumption can occur in physical or virtual form. In physical self-consumption, generation plants are directly connected to consumption loads within the private network of a building or condominium, with a single point of connection (POD) to the public grid, shared supply contract and internal metering infrastructure. In virtual self-consumption, each user maintains its own POD and can freely choose the electricity supplier; physical energy exchanges occur over the public grid, ending at each user's withdrawal points, with the grid operator responsible for measurement.

A REC is defined as an autonomous legal entity, typically established as an association or cooperative and without a profit-seeking purpose. Its members commit to produce, share and consume clean energy generated within the community by entities classified as producers or prosumers. Membership does not limit the rights of the individual final customer, who retains the freedom to choose the electricity supplier and the right to exit at any time, subject to the bylaws and constitutive agreements. A REC should comprise at least two members and at least two distinct connection points, one for consumption and one for production, with no upper limit on participants. Eligible participants include private citizens, small and medium enterprises, territorial authorities (municipalities, provinces, regions, metropolitan cities), local authorities, research and training bodies, religious and third-sector associations, or environmental-protection organizations, if participation is not their main economic activity. Members can assume the roles of consumer, producer or prosumer: consumers use shared energy to meet part of their needs, producers generate renewable energy for the community, and prosumers combine production and self-consumption, covering their own loads and sharing surplus with other members.

REC plants should be fed exclusively by renewable sources such as photovoltaic, wind, hydro, biomass and biogas, possibly combined with storage systems. Within the community, energy is shared among producers and consumers connected to the same primary substation, using the national distribution grid. This is called virtual energy sharing: the shared energy corresponds to the minimum between the total energy injected into the grid by renewable plants and the total energy withdrawn by members, thereby maximizing collective self-consumption and promoting a more efficient and sustainable power system. A key role in managing the Energy Community is played by the Referente (legal representative), who handles the technical and administrative management of the REC. This function can be performed by the community itself through its legal representative, by a producer or a final-customer member, or by an external third-party producer holding a certified UNI 11352 ESCO licence, provided that its plant contributes to the community configuration. The Referente coordinates plant operation, ensures correct accounting of energy flows and production/consumption data, and serves as contractual counterpart with the GSE, submitting incentive-access requests and receiving related invoices.

3.3 The Digital Platform MyCER

Italy's Legislative Decree 199/2021 and the Ministerial Decree n. 414/2023 have enabled new Collective Renewable Energy Self-Consumption Communities, which require not only aggregating individual plants and loads but also continuous coordination, monitoring and regulatory compliance. Building on this framework, Higeco Energy—a company specializing in industrial and PV-plant monitoring—has used its robust cloud infrastructure to create a user-friendly front-end for end users unfamiliar with PV and consumption monitoring.

MyCER platform has been developed as a robust, scalable tool that helps RECs actively participate in electricity markets. In both standard and white-label modes, MyCER centralizes data from all PODs, ensures traceability and compliance with GSE incentive rules, and optimizes energy flows via dedicated algorithms to maximize collective self-consumption and reduce imbalances. It offers transparent dashboards for administrators, referees and members, improving trust and engagement, and supports scalable governance across multiple configurations—Collective Renewable Energy Self-Consumption Communities, collective self-consumption groups and remote individual self-consumers—on a single platform. In this sense, MyCER translates complex regulations into simple operations, turning energy data into economic and environmental value for all participants. For the URB-REC project, MyCER will be used for simulation, exploring multiple scenarios to recommend, before any investment, the configuration closest to the Referente's preferences, thereby reducing the risk of errors and improving the expected return on future investments.

4 Case Study

4.1 Geographical Context and Spatial-planning Framework

The Municipality of Masullas, a little town (~1,000 inhabitants) in Oristano province, central-western Sardinia (Fig. 1), serves as a case study representative of most Sardinian municipalities (characterized by a population of about 1000 inhabitants) and due to the energy-transition initiatives carried out by the municipal administration, including the ongoing update of the Local Plan. Spanning 19 km2 with a trapezoidal shape bounded by ridges and Monte Arci slopes (altitudes 81–409 m), it features diverse microclimates, hills, agriculture, natural areas, a hydrographic network, and a temperate-Mediterranean climate. Located in Monte Arci Regional Park, it boasts biodiversity, prehistoric obsidian sites, and trails for sustainable tourism/education.

The economy relies on agro-pastoral-silvicultural systems, diversifying into olives, vineyards, and crops with sustainable/self-production focus. Land use: herbaceous crops (45.7%), tree crops/natural areas; built-up non-urban $<2\%$; forests minimal.

The compact urban center blends a preserved historic core with expansions along Via Nazionale, offering distributed services (greens, sports, squares, public/religious buildings) and high-quality environment. Key sites: "Sa Gloriosa" parish church, "Palazzo Salis" (town hall/museums), historic churches. Well-connected via SS131 road, Uras railway, and local roads despite peripheral position, Masullas harmonizes natural/agricultural/urban elements with cultural heritage, supporting a rural model of

environmental protection, innovation, and tourism—ideal for a Local Plan-driven energy goals. Masullas' Local Plan, adopted in 1999 and last generally amended in 2011, defines the ancient urban core perimeter. The 2008 Historic Centre Detailed Plan (Fig. 1) implements the Local Plan goals: conserving public spaces, protecting monuments/relational areas, enhancing pedestrian quality, and linking key sites. Its "PROS 4 Energy and Sustainability" project promotes bioclimatic retrofits, solar optimization, and energy-efficient PV/solar system requiring certification, non-obtrusive integration (e.g., roof-fitting panels on hidden slopes, photovoltaic tiles in the historic center), and banning rooftop boilers (attic/courtyard placement only) to preserve historic forms.

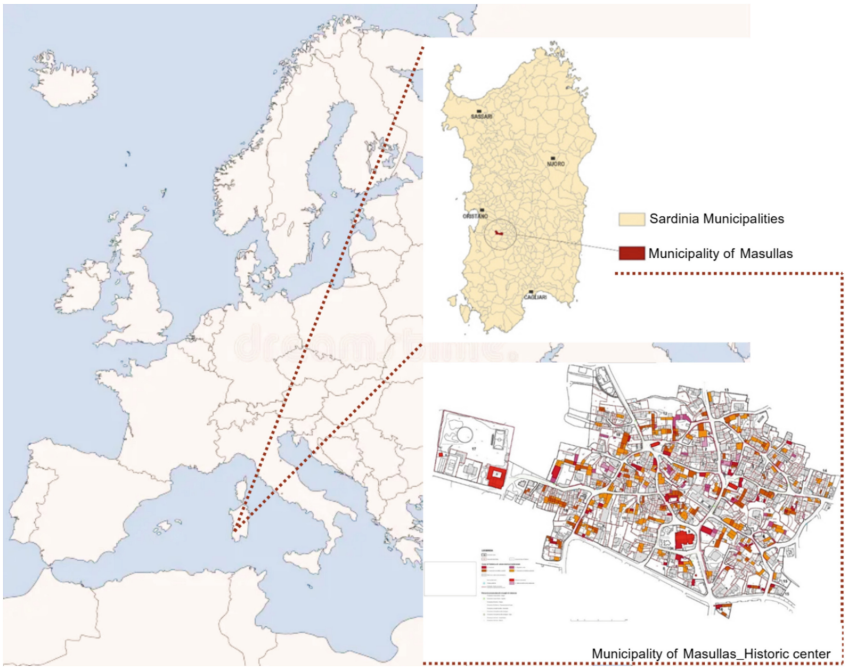


Fig. 1. The geographical context of the Municupality of Masullas (Sardinia, Italy), and an extract from the zoning table of the Historic Centre Detailed Plan.

4.2 Public Electricity Consumption

In accordance with the objectives of the URB-REC project, which aims to develop a REC model for the energy regeneration of the historic urban landscape, data on both public and residential electricity consumption were analyzed. Specifically, the local administration (the Municipality of Masullas) provided direct annual electricity consumption data for the main municipally owned properties used for public services in the historic center. The utilities studied include public buildings and spaces which can be divided into various categories: cultural and museum buildings, schools and sports facilities, and religious buildings.

In addition to the list of municipally owned properties and areas, the administration also provided the three PODs for public lighting systems, which were also studied.

To process the monthly and hourly electricity consumption data, necessary to characterize the demand profiles of the various users, historical withdrawal data provided by the grid operator for each POD were used. Consumption was further broken down into time slots F1, F2, and F3 to highlight load variations based on the consumption time slot.

Domestic electricity consumption, however, in the absence of direct measurements, is defined based on data provided by the ARERA portal to ensure a consistent and representative comparison.

The methodologies used to estimate monthly consumption, both for municipally owned properties and for residential users, allow user load profiles and identify peaks in demand.

5 Results

Consistently with the objectives of the URB-REC project, aimed at promoting electricity generation from renewable sources in an urban context by exploiting public real-estate assets, municipal properties in Masullas hosting the main public services in the historic center were selected, for which direct consumption data are available. The electrical supplies under analysis (2024) include cultural and museum buildings, educational and sport facilities, religious buildings, and the three PODs related to public lighting, which were also analyzed.

Consumption was analyzed on a monthly and hourly basis for each POD, reconstructing load profiles and identifying demand peaks. For some municipal buildings, missing data were estimated by correlating them with a reference building (e.g., the former Town Hall) equipped with a complete hourly profile; for domestic consumption, for which no direct measurements were available, a standard ARERA profile was applied and disaggregated by time band. The overall hourly profile was reconstructed by assigning the former Town Hall's profile to the various PODs and, where necessary, by applying specific profiles for public lighting and the municipal supply, integrating all hourly consumption values.

MATLAB was used to simulate energy flows for different photovoltaic capacities, starting from the hourly consumption data of municipal supplies and the solar irradiation data for the province of Oristano. The results indicate higher gains in the first 100 kWp, with load coverage approaching 35% above 500 kWp, leaving about 65% of consumption uncovered. To increase coverage, energy storage systems are therefore required (Fig. 2).

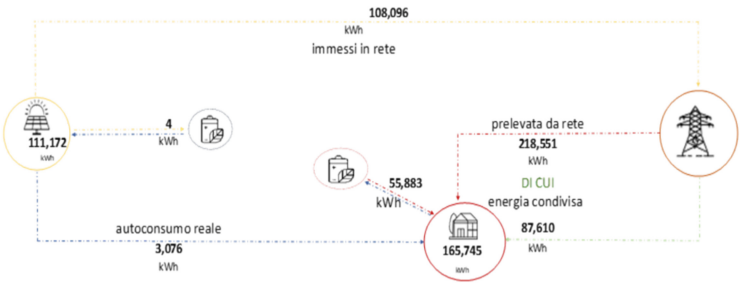


Fig. 2. Energy flows of the configuration with public administration supplies.

The available surface on buildings and public areas suitable technologically and compatible from an urban-planning perspective for photovoltaic installations was mapped, identifying, in agreement with the Municipality, areas potentially usable for a REC and preparing descriptive sheets for each one, including the following information: type of asset, location and geographical context, urban-planning classification, installation surface type, overall capture surface.

Finally, MyCER simulator was used to verify a Collective Renewable Energy Self-Consumption Communities configuration with a 60 kWp photovoltaic system and a 140 kWh storage capacity, capable of guaranteeing about 55% self-sufficiency. The hourly annual simulation shows that more than half of the community's consumption is locally covered, with flows of feed-into the grid, self-consumption and energy sharing that, combined with storage, increase coverage of the demand-side users. The cash flows, discounted over 20 years, show a progressive reduction of revenues over time due to discounting and the gradual decline in photovoltaic system efficiency, as well as a reduction in revenues from grid-injection and shared energy. The entire incentive, net of management costs, is allocated to the producing supplies (the Municipality), while the payback period of the investment is calculated considering both the photovoltaic installation alone and the overall Collective Renewable Energy Self-Consumption Communities system.

6 Conclusions

Renewable Energy Communities (RECs) emerge as a tool that links conservation-oriented urban planning with energy and mobility planning in a unified, systemic and interdisciplinary way, supporting a new model of sustainable historic city. They also help limit visually intrusive technical installations to protect landscape identity, but any new visible equipment must be carefully assessed for its visual and cultural impact to avoid disrupting the environment's formal balance and authenticity.

Growing demands for efficiency, connectivity, safety and smart services have led to uncoordinated, fragmented installations. To reduce this, a multidisciplinary approach is needed, combining regulatory rules, design guidelines and monitoring that specify materials, colours, positioning and size of plants, while favouring integrated or "camouflaged" technologies and coordinating actors to avoid overlapping interventions.

The URB-REC project develops this idea as a multidisciplinary proof-of-concept that integrates urban, energy and mobility planning to promote the sustainable conservation of historic centers through RECs, especially in Sardinia's inner areas. The model treats RECs as energy-transition ecosystems based on coordinated participation, mainly led by public institutions, which protect historic public buildings while decarbonizing through collective renewable-energy production. The project validated the legal framework (RED II, IEM, Italian Decrees 199/2021 and 210/2021), identified simplified procedures for PV in and near protected areas, and applied this approach in Masullas. Using the MyCER platform, the project also enabled savings and incentive distribution, user-level monitoring of consumption and generation, integration with storage and e-mobility infrastructures, and context-specific, white-label adaptation. Key benefits are collaborative, equity-oriented models that help combat energy poverty.

The URB-REC project and the Masullas case show this is technically and economically feasible, but its success depends on the political will to implement these changes. As a matter of fact, the model is viable and replicable but requires further regional-scale experiments to streamline landscape authorizations and spread digital management tools for RECs, in turn making them strategic instruments for urban sustainability and resilience in historic areas.

To turn RECs into transformative tools in historic centers, the study proposes a move from incentive-centered to governance-centered policies, building capacity through facilitation services, legal-technical support and training. It also calls for simplifying complex regulations, fostering community building and social-capital development, and embedding energy-justice principles that include vulnerable groups excluded from installing PV on their own buildings. Communication should stress collective, relational and environmental benefits rather than only individual savings, and policies should strengthen municipal leadership, support inter-municipal cooperation and invest in long-term institutional development via multi-year funding and adaptive learning.

Acknowledgments. This study was funded by the research project URB-REC "Managing the Historic Urban Landscape: Renewable Energy Communities for Sustainable Urban Conservation" – UNICA PoC Call (DR 268/2025 of 12/03/2025) for the development of Proofs of Concept (PoC) by researchers aimed at promoting technological innovation and the energy transition - within the framework of Spoke 7 of the e.INS project - Ecosystem of Innovation for Next generation Sardinia - MUR Code: ECS00000038 - CUP F53C22000430001. Principal Investigator (PI): Mara Ladu. Research Group: Ginevra Balletto, Stefano Matta, Emilio Ghiani, Francesco Puddu, Viola Gioi, Alessandra Milesi, Federica Isola and Corrado Zoppi (University of Cagliari), and HIGECO ENERGY srl.

Author Contributions. G.B. and M.L. wrote Sects. 1, 3.1 and 6; G.B. wrote Sect. 2.1; F.I. wrote Sect. 2.2; E.G. wrote Sect. 2.3, 3.2 and 4.2; M.L. wrote Sect. 3.3; F.I. and A.M. wrote Sect. 4.1; G.B., M.L., F.I., E.G., F.I. and A.M. wrote Sect. 5. The authors have read and agreed to the published version of the manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fotiou, T., Fragkos, P., Zisarou, E.: Decarbonising the EU buildings| model-based insights from European countries. Climate **12**(6), 85 (2024)
2. Tarpani, E., et al.: On renewable energy community implementation in historic cities: a city-scale validated model. Energy Build. **338**, 115709 (2025)
3. Balletto, G., Ladu, M., Milesi, A.: Città circolare e transizione energetica, tra tutela e valorizzazione dei centri storici. In: Gerundo, R. (eds.) Città e piani del rischio energetico e alimentare, pp. 27–40. Edizioni Scientifiche Italiane spa (2023)
4. Jawad Ul Hassan, M., Belloni, E., Faba, A., Cardelli, E.: Energy communities, renewables, and electric mobility in the Italian scenario: opportunities and limitations in historic town centers. Energies **18**(22), (19961073) (2025)
5. Mangano, G.: Renewable energy communities: enabling technologies and regenerative models for the green and digital transition in the inner areas: investigation through case studies for the experimentation of living lab in the grecanica area. In: In: Bevilacqua, C., Balland, PA., Kakderi, C., Provenzano, V. (eds.) International Symposium: New Metropolitan Perspectives, pp. 309–331. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-34211-0_15
6. Leone, F., Zoppi, C.: La pianificazione dei Centri storici in Sardegna: una visione strategica per la redazione dei piani particolareggiati nel contesto del Piano paesaggistico regionale. Planum **30**(1), 1016–1024 (2014)
7. Isola, F., Zoppi, C., Pira, C. (eds.): Centri storici e spazio urbano – L'esperienza regionale della Sardegna. FrancoAngeli, Milano (2018)
8. Ladu, M.: Conservation and regeneration for a sustainable and circular city. In: Gervasi, O., et al. (eds.) International Conference on Computational Science and its Applications. ICCSA 2023. LNCS, vol. 14111, pp. 77–88. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-37126-4_6
9. De Santoli, L.: Guidelines on energy efficiency of cultural heritage. Energy Build. **86**, 534–540 (2015)
10. Balletto, G., Ladu, M., Camerin, F., Ghiani, E., Torriti, J.: More circular city in the energy and ecological transition: a methodological approach to sustainable urban regeneration. Sustainability **14**(22), 14995 (2022)
11. Balletto, G., Ladu, M., Borruso, G.: Transizione demografica e urbanizzazione: alcuni spunti di riflessione di un processo in corso. In: Gerundo, R. (eds.) Città e piani del rischio demografico, pp. 27–43. Edizioni Scientifiche Italiane spa (2024)
12. Sinatra, M., Balletto, G.: Criteri ambientali minimi nella giustizia ambientale. In: XXV Conferenza Nazionale siu Società Italiana degli Urbanisti Transizioni, Giustizia Spaziale e Progetto di Territorio, Planum, vol. 10, pp. 139–144 (2024)
13. UNID Professional homepage. https://www.unidprofessional.com/rigenerazione-urbanistica-quadro-normativo/. Accessed 20 Sep 2025
14. Malcevschi, S.: Rischi per la transizione ecologica corrente e nuove sfide per ecosistema e territorio. Scienze del Territorio **10**(2), 29–37 (2022)
15. Direttiva (UE) 2018/2001 del Parlamento Europeo e del Consiglio dell'11 dicembre 2018 sulla promozione dell'uso dell'energia da fonti rinnovabili. https://eur-lex.europa.eu/legalcontent/IT/TXT/PDF/?uri=CELEX:32018L2001
16. Strategia Energetica Nazionale (2017). https://www.mimit.gov.it/images/stories/documenti/Testo-integrale-SEN-2017.pdf. Accessed 20 Sep 2025
17. ENEA magazine homepage. https://www.eai.enea.it/component/jdownloads/?task=download.send&id=378&catid=14&Itemid=101
18. Il piano nazionale di ripresa e resilienza. https://temi.camera.it/leg19/pnrr.html

19. Isola, F., Leone, F., Lai, S.: Efficienza energetica e pianificazione dei centri storici: alcune esperienze dalla regione Sardegna. Archivio di studi urbani e regionali **131**(supplemento), 118–142 (2021)
20. Pellegrini, P., Micelli, E.: Paradoxes of the Italian historic centres between underutilisation and planning policies for sustainability. Sustainability **11**(9), 2614 (2019)
21. MiBACT – Ministero dei beni e delle attività culturali e del turismo Linee di indirizzo per il miglioramento dell'efficienza energetica nel patrimonio culturale. Architettura, centri e nuclei storici ed urbani (2015). https:// soprintendenza. pdve. beniculturali.it/ wp-content/. Accessed 7 May 2026
22. Trevisan, R., Ladu, M., Ghiani, E., Balletto, G.: Achieving net zero condominiums through energy community sharing. Sustainability **16**(5), 2076 (2024)
23. Ahmed, S., Ali, A., D'angola, A.: A review of renewable energy communities: concepts, scope, progress, challenges, and recommendations. Sustainability **16**(5), 1749 (2024)
24. Zulianello, M., Sala, F., Coletta, G., Armanasco, F.: Le Comunità Energetiche in Italia. Note per il coinvolgimento dei cittadini nella transizione energetica. RSE, Editrice Alkes per collana RSEview (2021)

How Do Local and Regional Strategies Address Blue-Green Infrastructure? the Case of Slovakia

Ondrej Mitaľ, Michaela Lukačínová, and Silvia Ručinská(✉)

Department of Public Policy and Theory of Public Administration, Faculty of Public Administration, Pavol Jozef Šafárik University in Košice, 04011 Košice, Slovak Republic
{ondrej.mital,michaela.lukacinova,silvia.rucinska}@upjs.sk

Abstract. The study investigates blue-green infrastructure, the essential factor of sustainable urban development. Blue-green infrastructure helps cities and regions, among other things, to generate benefits in providing services, strengthening resilience and solving environmental issues and crises. The study uses literature review and content analysis to examine blue-green infrastructure and the contexts in which it is used. Regarding the conditions of the Slovak Republic, altogether eight strategic documents of regional cities and eight strategic documents of regions were analysed, focusing on partial components of blue-green infrastructure. Based on the literature review and content analysis of local and regional strategic documents in selected self-government units, seven complementary components of the blue-green infrastructure have been identified, green infrastructure, administrative roles, green-blue infrastructure services, green-blue infrastructure benefits, climate change adaptation, community impact, and blue infrastructure. This study's primary benefit lies in its contribution to the ongoing discussion about the sophistication of blue-green infrastructure, which supports resilient urban development and the sustainable growth of local and regional communities.

Keywords: blue-green infrastructure · blue-green infrastructure benefits · green infrastructure · blue infrastructure · strategic documents · Slovakia

1 Introduction

Environmental challenges as a result of climate change and green transition require responses from public administration at different levels of government. Responses of states, regions and cities involve different elements of blue and green infrastructure, which complement traditional grey infrastructure functions in urban areas. As a result of long-term development, the combination and synergic effect provided by a combination of grey, green and blue infrastructure leads to the emergence of green-blue infrastructure elements and solutions. Despite the rising integration of the term blue-green infrastructure into different types of policies and strategies, this phenomenon is still relatively unexplored, especially in terms of its components and evaluation (de Rijke et al. 2025; Molné et al. 2023).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 592–609, 2027.
https://doi.org/10.1007/978-3-032-30521-3_37

The study aims to contribute to the intensive scientific discussion on the importance of blue-green infrastructure in its theoretical and practical aspects. The study aims to realise preliminary research and investigate the blue-green infrastructure agenda in the Slovak Republic. The study uses qualitative and quantitative aspects of content analysis to examine blue-green infrastructure verbal data content included in strategic documents adopted at Slovakia's regional and local levels. The research sample consists of eight self-governing regions and eight regional cities in the Slovak Republic, more precisely, mandatory strategies called Plans of economic and social development in selected regions and cities.

2 From Grey to Blue-Green Infrastructure

Traditional engineering approaches based on grey infrastructure are not able to meet all the challenges of sustainability associated with growing urbanisation, posed by increasing urban population and consumption (Boogard 2024; Perrelet et al. 2024; Pinto et al. 2023; de Oliveira et al. 2022). Grey infrastructure is a usual and standard man-made infrastructure or solution designed to provide individuals and businesses with specific services and its role is mainly functional support for urbanised areas (Wang and Banzhaf 2018; Liquete et al. 2015). Grey infrastructure alone is not capable of addressing contemporary environmental trends and challenges.

Climate change is increasing the magnitude and frequency of numerous environmental challenges, mainly floods, storms, and landslides, making adaptive solutions crucial (Capobianco et al. 2024; Das 2024). In the last decade, public administration has continuously shifted from grey to blue-green infrastructure (European Union, 2024). In the last decade, we can identify a growing emphasis on the functions and benefits of blue-green infrastructure at local, regional, national and European governance levels. The term blue-green infrastructure originally consisted of green infrastructure and blue infrastructure. Compared to traditional grey infrastructure, blue-green infrastructure can provide multiple functions to urban areas.

2.1 Green Infrastructure and Blue Infrastructure

Green infrastructure has become an essential part of policies and strategies at different levels of government worldwide, mainly due to the increasing environmental and natural issues and disasters. The successful green transition depends, to a large extent, on the ability of various government levels to deliver sustainable infrastructure elements (OECD, 2022). The importance of green transition has resulted in integrating green infrastructure into urban planning to address green transition in our society (Vitulano 2024). This concept emphasises integrating green solutions with other grey infrastructure systems, creating new elements and redesigning existing and traditional urban solutions (Grabowski et al. 2022). Green infrastructure should be used by urban and building designers to achieve wider environmental goals (Pereira et al. 2023).

European Commission sees green infrastructure as a strategically planned network of natural and semi-natural areas with other environmental features designed and managed to deliver a wide range of ecosystem services (European Commission 2021).

From a practical point of view, green infrastructure helps states, regions, cities and communities utilise environmental, social and economic development with a focus on sustainability (Korkou et al. 2023; Ying et al. 2022; Lennon 2014). Based on systematic research, Ying et al. (2022) created a spatial scale composed of (1) region/macroscale (green corridors, green patches, wildlife migration corridors, green countryside), (2) city/mesoscale (green alley and street, public park, urban wetland, urban forest, urban waterfront, urban greenway), and (3) site/micro-scale (rain garden, grass channel, retention ponds, permeable pavement, ecological parking, green roof).

Building green infrastructure is not just a goal but a powerful catalyst for social, environmental and economic benefits. To be more precise, green infrastructure holds the potential to create social benefits (encouraging healthy and sustainable lifestyles, supporting urban communities, investing in liveable spaces), environmental benefits (enabling long-term climate resilience, making smart and connected landscapes, fostering urban biodiversity), and economic benefits (integrating urban food, renewing urban spaces, allowing energy and resource efficiency) (ARUP 2014).

With the rising importance of green infrastructure, blue infrastructure is also being implemented in policies and strategies. The strong emphasis on blue infrastructure is ultimately the result of decades of economic development, urbanisation, as well as the degradation and conversion of natural land into areas used for human activities (Iojă et al. 2021).

States, regions, cities and communities aim to utilise blue infrastructure to mitigate urban warming, and extreme heat has become a widely concerned issue for promoting cooler cities (Cheng et al. 2023). Blue infrastructure is an important heat mitigation strategy that alleviates thermal discomfort in urban infrastructure (Cheng et al. 2023; Lin et al. 2020).

Blue infrastructure refers to the network of water bodies and water management systems in urban and rural areas. It refers to natural bodies such as rivers, lakes, ponds, wetlands, other water features, and engineered structures, such as drainage systems, canals and flood control mechanisms (Severin and Michalíková, 2024). According to the character of partial elements, the design and construction of blue infrastructure require sophisticated, in-depth observation of the natural systems and incorporation of desired outcomes of the interventions (Bugnot et al. 2018).

Due to its predominant use in the field of drainage and water management, practitioners and researchers from various disciplines are increasingly using the term ‘blue infrastructure’ to refer to water-related surfaces such as wetlands, rivers and ponds, or to strategies employed for flood control (de Oliveira et al. 2022; Andersson 2018; Pauleit et al. 2017).

Blue infrastructure represents an integral part of environmental public policies at different levels of government. However, this term is mainly used as a part of or in combination with green infrastructure. Thus, the most used term in contemporary times is blue-green infrastructure.

In summary, green infrastructure refers to natural and semi-natural elements, such as parks, gardens, urban forests, and green roofs, while blue infrastructure encompasses water bodies, waterfronts, and aquatic features (Das 2024). Green infrastructure and blue infrastructure elements can fulfil multiple purposes. The key is not only the creation of

these elements in the territory but also the evaluation of the quality of these elements, which are intended to achieve the desired benefits (Wang and Banzhaf 2018). Current trends and challenges forced states, regions and cities to combine the benefits of green infrastructure and blue infrastructure in its multidimensional context.

2.2 Blue-Green Infrastructure

The connection of green infrastructure and blue infrastructure elements has a growing impact on contemporary environmental policies at national, regional and local levels of governance. The conceptual approach to green and blue infrastructure elements is also gaining interest at the European level. Key stakeholders perceive the importance and the potential of integrating green and blue elements into blue-green infrastructure.

Integrating the green (soft areas, plants and trees) with the blue (watercourses, ponds, lakes and storm drainage) elements makes our urban spaces more resilient, pleasant and healthy places to live, work and play (GreenBlue Urban 2025). These integrated solutions resulting from green (land) and blue (water) spaces can improve environmental conditions and, therefore, citizens' health and quality of life (European Commission 2021).

Blue-green infrastructure refers to an interconnected network of natural and semi-natural elements that provide multiple environmental, social, and economic benefits, emphasising the value that nature provides to human society and also helps to avoid relying on infrastructure that is expensive to build when nature can often provide cheaper, more durable solutions (Severin and Michalíková, 2024). Put differently, blue-green infrastructure initiatives can be implemented to provide cost-effective solutions to urban challenges and hence advance the achievement of many Sustainable Development Goals (de Oliveira et al. 2022). In this regard, green-blue infrastructure combines the functions and services of green infrastructure and blue infrastructure to utilise their impact on environmental issues and sustainable development.

Blue-green infrastructure is currently very often grounded in nature-based solutions. A stronger integration of nature-based solutions into nested green and blue infrastructure planning could be a promising way forward in urban and landscape planning (Langemeyer and Baró 2021). Nature-based solutions are inspired and supported by nature, are cost-effective, simultaneously provide environmental, social and economic benefits, and help build resilience. Such solutions bring more diverse nature and natural features and processes into cities, landscapes and seascapes, through locally adapted, resource-efficient and systemic interventions (European Commission 2025). Nature-based solutions are actions to protect, sustainably manage or restore natural ecosystems that address societal challenges such as climate change, human health, food and water security, and disaster risk reduction effectively and adaptively, simultaneously providing human well-being and biodiversity benefits, for example (World Bank Group 2022). Using nature-based solutions can be more cost-effective and create additional co-benefits than relying solely on conventional infrastructure, because overall capital costs are generally lower for nature-based solutions than for conventional infrastructure, although maintenance costs can be higher (OECD, 2023). Typical examples include soil system degradation management, watershed protection, and forest protection and restoration.

Analysis of 850 peer-reviewed publications revealed that green infrastructure and blue infrastructure is associated with important ecosystem services such as local temperature regulation, climate change adaptation, recreation and aesthetic appreciation, benefits of parks and other different uses of green spaces, while some ecosystem services are still developing and are still relatively unexplored (Veerkamp et al. 2021).

From the practical point of view, blue-green infrastructure elements typically combine green infrastructure, such as parks, green roofs, and urban forests, with blue infrastructure, which includes water bodies like rivers, lakes, and wetlands (Severin and Michalíková, 2024). Blue-green infrastructure elements can be applied on different levels of government and implemented alone or associated with others (Meng et al. 2023). With regard to its scope and interdependence, blue-green infrastructure ideally refers to a strategically planned network of natural and semi-natural areas and tools with other environmental features (European Commission 2018).

A wide range of environmental services provided by blue-green infrastructure elements promoted blue-green infrastructure to the multifunctional concept (Cook et al. 2024; Romańczak et al. 2024; Korkou et al. 2023; Wang and Banzhaf 2018; Bears 2018; Breuste et al. 2015). The multidimensionality of blue-green infrastructure also leads to various benefits perceived by relevant stakeholders living in regions and cities.

2.3 Blue-Green Infrastructure Benefits

The multidimensionality of blue-green infrastructure directly indicates that blue and green elements in the territories generate a different and wide range of benefits. Generally speaking, blue-green infrastructure positively impacts the urban population and fundamental living functions (Azadgoleh et al. 2025; Haase 2015).

The practical benefits of blue-green infrastructure, among the most important ones, are usually expressed through flood mitigation, biodiversity enhancement, improved air quality, climate resilience, recreational spaces, aesthetic value, and economic benefits (Severin and Michalíková, 2024). Blue-green infrastructure also has a significant direct and secondary influence on various environmental mechanisms in the territory, such as vegetation, soil system, water transfers, or wind direction (Jones et al. 2024).

Blue-green infrastructure represents an important attribute of sustainable development; as two sides of the same coin, it has the potential to solve a number of problems and, at the same time, ensure benefits.

The flow of benefits from GBI to inhabitants and relevant local stakeholders is mostly mediated by three filters: framework, infrastructures, institutions and perceptions (Kronenberg et al. 2021). The combination of these three filters is generally outlined and defined in different types of policies, strategies, documents and action plans. Policies and strategies connect visions and goals with practical elements of blue-green infrastructure, as well as the services and benefits they are built and maintained for.

Benefits generated by blue-green infrastructure must be systematically planned and purposefully coordinated. Put differently, blue-green infrastructure elements must be built and maintained in accordance with policies and strategies. Systematisation of blue-green infrastructure benefits through policies and strategies can be perceived as an important part of sustainable development, especially thanks to systematic urban

Table 1. Benefits of blue-green infrastructure

Authors	Benefits of blue-green infrastructure
Fenoglio et al., 2024; Perrelet et al. 2024	Biodiversity
Benoit et al. 2025; Perrelet et al. 2024; Tate et al. 2024; Choi et al. 2021; Rouse and Bunster-Ossa 2018; Cole et al, 2017; Deak and Bucht 2011	Climate change adaptation
Ferreira et al. 2024; Hanna et al. 2024; Sun et al. 2024; Vogel 2024	Climate change mitigation
European Commission 2025; Bray et al. 2024; Iojă et al. 2021; Haase, 2015; Kernaghan and Sturgeon 2024; Lõhmus and Balbus 2015	Health and liveability
European Commission 2025; Bray et al. 2024; Das 2024; Kim et al. 2024; Pinto et al. 2023	Quality of life and wellbeing
Kim et al. 2024; Jones et al. 2024; van Renterghem et al. 2012	Reduction of pollution
Ambily et al. 2024; Boogard 2024; Monaco 2024; Ojo 2024; Anderson et al. 2021	Resilience of the ecosystem
Monaco 2024; Tetrytskyi 2024; Cole et al. 2017	Sustainable development
Azadgoleh et al. 2025; Monaco 2024; Shi 2020	Social justice, inequality and social marginalisation

Source: authors

governance and the involvement of relevant local stakeholders (Scott and Hislop 2024; Pinto et al. 2023; Johns 2019).

Blue-green infrastructure represents a narrative integrated into different sectoral policies in the past 10 years (Chatzimentor et al. 2020; Hislop et al. 2019; Artmann et al. 2019; Lennon, 2014). Containing green-blue infrastructure in policies and strategies has become common practice (Shi 2020).

2.4 Blue-Green Infrastructure - Current Research Gap

Based on the abovementioned, building and maintaining blue-green infrastructure can be perceived as a trend among the regions and cities. This trend focusing on environmental challenges can be perceived as intentional steering of urban development that shall secure required climate change actions and desired benefits for citizens.

Despite unquestionable blue-green infrastructure benefits, opinions continue to emerge, pointing out that this phenomenon still needs deeper research. It is still perceived as a relatively new term in urban planning and governance, especially in its practical meaning (Marot et al. 2024; Scott and Hislop 2024; Simon et al. 2021; Veerkamp et al. 2021).

Based on the research conducted so far, there is an evident need to define and clarify what we understand by these terms. (Grabowski et al. 2022; Matsler et al. 2021).

Theoreticians also mention problematic assessment, the vagueness of the term GBI and its low impact on urban development (de Macedo et al. 2021). These doubts and uncertainties might lead to confusion, unintended consequences or unmet promises (Matsler et al. 2021). However, all these similar opinions can be countered by the idea that the concept of blue-green infrastructure was created as flexible, robust and adaptable to specific needs (Matsler et al. 2021; Garmendia et al. 2016).

3 Methods

The study aims to realise preliminary research and investigate the blue-green infrastructure agenda in the Slovak Republic. The study uses qualitative and quantitative aspects of content analysis to examine blue-green infrastructure verbal data content included in strategic documents adopted at Slovakia's regional and local levels. Ahead of this analysis, the study theoretically examines the phenomenon of blue-green infrastructure and its importance and benefits.

The study aims to answer research questions: (1) What is the frequency of use of the terminology of the blue-green infrastructure in the selected strategic documents adopted by selected regional and local self-governing units in the Slovak Republic? – What helped to examine whether the selected phenomenon is gaining serious attention from selected self-governing units? (2) In which contexts and through which elements do selected regions and cities use the term blue-green infrastructure? – to identify the decomposed meaning of blue-green infrastructure elements in practice; (3) What is the consistency and interconnectedness of strategies at the regional and local level? – to identify whether regional and local levels of self-government pay attention to similar aspects of blue-green infrastructure.

The research sample consists of eight self-governing regions and eight regional cities in the Slovak Republic. The research sample was selected intentionally; thus, the study covers all eight self-governing regions and eight regional cities, the selection of which results from their significance in the Slovak conditions. Regional cities provide public services for a wider territory. These eight cities are also the headquarters of the self-governing regions´ offices. The importance of eight regional cities is also underlined by the fact that 20,96% of the population in the Slovak Republic lives in regional cities. The status of a regional city is not associated with any special or unique legal status or competencies. The main difference between cities and other units is determined by the fact that cities realise key activities focused on the production of public services both for their residents and residents from the nearest territory. Slovakia's regional cities have even more financial, material, and human resources than small municipalities.

The content analysis focused on the frequency of blue-green infrastructure usage and on interpreting and understanding the context of examined phenomena in adopted strategic documents by selected self-governing units. The study analyses mandatory strategic documents that must be adopted by self-government units, mostly called Plans of development or Plans of economic and social development. Figure 1 shows the validity of the analysed regional and local strategies.

Content analysis was used to get an in-depth view of relevant aspects of blue-green infrastructure. The results of the content analysis identified the most frequently used

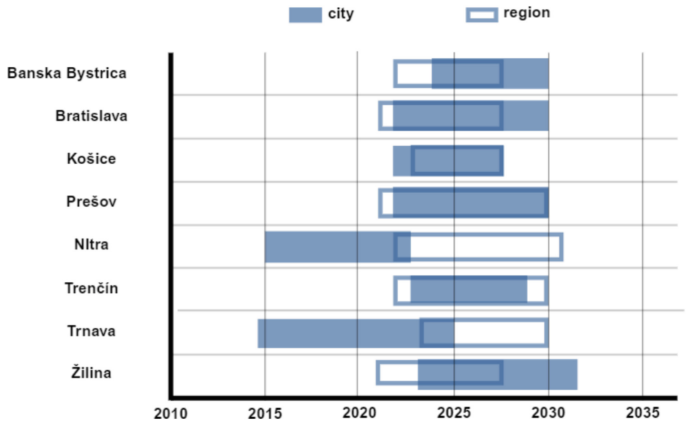


Fig. 1. Plans of Economic and Social Development in selected regions and cities – adoption and validity: Source: authors

terms in connection with blue-green infrastructure, which were subsequently grouped into clusters based on their content.

Theoretical analysis was used to analyse the contemporary state of research focusing on the concept of blue-green infrastructure. A literature review was used to integrate findings and perspectives from empirical studies. Previous research from different scientific disciplines was investigated to demonstrate the contemporary state of knowledge in the sphere of blue-green infrastructure, an important part of this part of ecosystems and the benefits of blue-green infrastructure.

4 Findings and Discussion

The term blue-green infrastructure is included multiple times in almost all analysed strategies, except one city and two regions. Put differently, the term blue-green infrastructure is used 71 times, 49 times in cities´ strategies and 22 times in regions´ strategies. Other key decomposed terms are green-blue areas, mentioned 72 times and green-blue solutions mentioned 15 times. The decomposition of blue-green infrastructure on green and blue aspects reveals a difference between the occurrence rate of green infrastructure (75 times) and blue infrastructure (23 times).

Strategies of selected regions and cities define blue-green infrastructure in different intensities and contexts. Altogether 919 terms connected to blue-green infrastructure were identified, more precisely 493 at the local and 426 at the regional level. The following clusters were identified based on qualitative content analysis of selected strategies adopted by regions and cities in the Slovak Republic (Fig. 2).

4.1 Blue-Green Infrastructure Services and Benefits

The cluster of blue-green infrastructure services includes services for the ecosystem with elements of blue-green infrastructure. By service, we mean an action that has a targeted

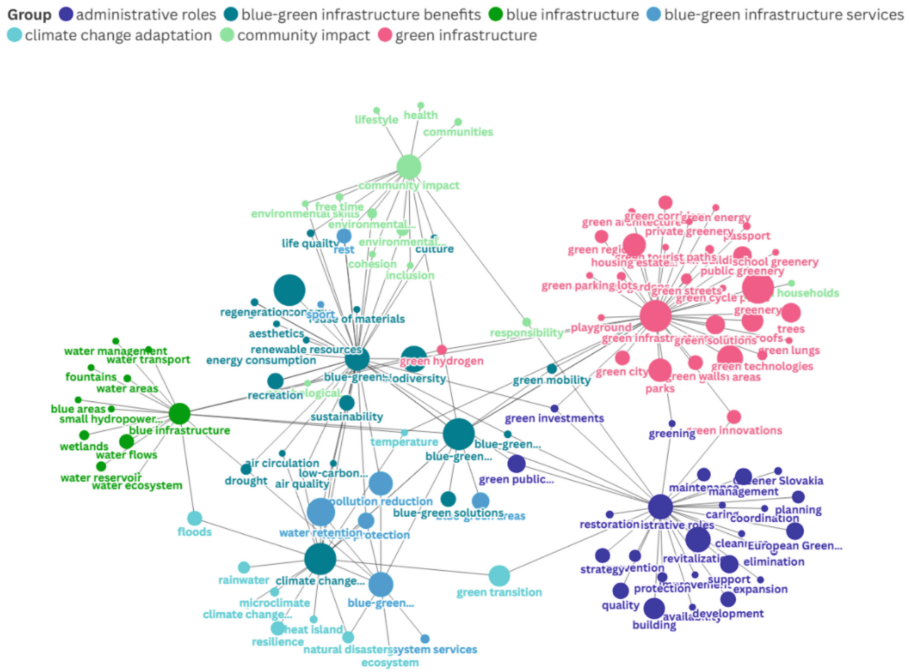


Fig. 2. Clusters of blue-green infrastructure contexts: Source: authors

and positive impact on the ecosystem or its specific parts, such as other parts or relevant local or regional stakeholders. In this regard, services should be perceived as not random but rather systematic and planned efforts of self-government units to change the current state of the ecosystem.

The primary benefits of blue-green infrastructure mentioned in the analysed documents respond to the need to reflect on solving current trends, challenges and issues in contemporary ecosystems. Climate change adaptation is the most frequently used term, followed by key activities of regions and cities that help to reduce the negative impact of climate change, mainly flood protection, water retention, pollution reduction and drought reduction. At the same time, these conscious activities help regions and cities prevent and overcome the negative impact of natural disasters.

4.2 Blue-Green Infrastructure Benefits

The cluster blue-green infrastructure benefits groups of words connected with helpful or good effect elements of blue-green infrastructure. Priority is given to benefits that actors directly intend as an improvement to the current state. Benefits that may not be directly intended by a specific solution or element of blue-green infrastructure are also included here.

Direct benefits resulting from blue-green infrastructure mainly include climate change adaptation, flood protection, water retention, and pollution reduction. Thus, these benefits directly contribute to preparing and solving the negative connotation of

green challenges and issues. Positive benefits are mentioned mainly in the context of biodiversity, air quality, and air circulation.

Subsequently, regions and cities perceive secondary benefits of blue-green infrastructure, mainly culture, aesthetics, recreation, sport, regeneration, rest, and free-time spending. The engagement of relevant communities must emphasise these benefits. Achievement of these benefits in regions and cities leads, based on the content of the documents, to better environmental awareness, effective environmental education and strengthening of environmental skills, as well as better inclusion and cohesion.

Selected documents see the benefits of blue-green infrastructure in the context of sustainability through lower energy consumption, reuse of materials and renewable resources.

4.3 Green Infrastructure

Another cluster of green infrastructure includes elements that can act independently due to their green character. These are elements that, due to their character and together with other elements and solutions, can contribute to the synergic effect of blue-green infrastructure.

The most frequently used terms in the context of green infrastructure are green areas, green solutions and green infrastructure. Based on the selected research sample, important terms are also green city and green region. Green buildings and green architecture are also terms that have a wider meaning for regions and cities, mainly in connection to green roofs and green walls.

An important topic of the analysed documents relates to green mobility, mainly in close connection to green streets, green corridors, green tourist paths and green cycle paths. Consideration of green infrastructure requires the use of greenery, while analysed documents use this term in different meanings, mainly public greenery, private greenery, housing estate greenery, and school greenery. Greenery is also indirectly linked to parks, trees and community gardens, which represent other important elements of green infrastructure. Subsequently, the term trees is also used with the word passport, which helps cities record all necessary information about trees in their territory. Analysed documents also, to a lesser extent, use another term, such as biodiversity, green technology, temperature, playground or green lungs.

4.4 Blue Infrastructure

Blue infrastructure includes elements that can act independently due to their blue or water character. These are elements that, due to their character, together with other components and solutions, can contribute to the synergistic effect of blue-green infrastructure.

The most frequently used terms in the context of blue infrastructure are blue areas, blue solutions and blue infrastructure. The following words are often used within the blue infrastructure context: biodiversity, water flows and water management. Blue infrastructure elements are discussed mainly through water reservoirs, fountains or small hydropower plants. At the same time, blue infrastructure is used in the context of negative phenomena that are used to prevent, mitigate or eliminate their negative impacts, mainly floods, rainwater and drought.

4.5 Climate Change Adaptation

Climate change adaptation includes areas focused on measures related to climate change. These are primarily activities aimed at adapting and mitigating contradictory impacts and creating elements of blue-green infrastructure to create positive impacts.

Selected regions and cities often connect climate change to broader terms, significantly impacting society and the economy. In this regard, documents discuss the importance of climate change adaptation in strengthening resilience and reflecting the green transition. Analysing strategic documents also puts pressure on natural disasters, floods, flood protection, water retention, and rainwater collection. Adaptation to climate change is also discussed with a focus on temperature, microclimate, air circulation, heat islands and low-carbon solutions.

4.6 Community Impacts

Community impacts are related to secondary effects that flow to different communities in an area. Positive impact depends not directly on a specific element but rather on the setting and relationships in a given community. Put differently, impacts may differ in different communities and blue-green infrastructure elements, as well as in regions and cities.

Selected self-government units declare the importance of communities through the analysed documents, mainly through the frequency of the word community, and also thanks to the words cohesion, inclusion and responsibility. Blue-green infrastructure is also expressed with its need to improve environmental awareness, education, and skills. This tendency suggests that regions and cities understand the importance of well-informed and educated citizens.

Blue-green infrastructure is subsequently mentioned in its impact on free-time spending, citizens resting, culture, sport, and also a positive impact on life quality and health.

4.7 Administrative Roles

Cluster administrative roles group operational and systematic activities of selected regions and cities related to blue-green infrastructure.

Among other issues, the administrative roles cluster groups strategic and systematic activities of regions and cities, especially thanks to the reflection of strategical documents European Green Deal and the national document Greener Slovakia.

Other partial activities focus attention on planning, development, management, protection, coordination, building, restoration, expansion and support. Greening and caring can be perceived as very specific roles connected to the existence of blue-green infrastructure in the territory. Subsequently, green public procurement and green investments can also be perceived as a direct result of green transition, which is also discussed in the context of the analysed strategic documents.

4.8 Compliance and Interconnectedness of Regional and Local Strategies

Regional and local levels of self-government should generally collaborate to ensure the development of certain territories. The conceptual basis for sustainable development and strengthening of blue-green infrastructure should be perceived as an important part of short-term and long-term success. The main reason is that environmental and climate change challenges cannot be limited to a small territory. These challenges do not respect the administrative boundaries of regions and cities. In this regard, compliance of local development strategies with regional local strategies is not just a legal obligation, but also a rational consequence of environmental challenges character, including blue-green infrastructure issues.

Figure 3 shows the 15 most used terms connected to blue-green infrastructure identified during the analysis of the strategic documents of regions and selected cities in the Slovak Republic.

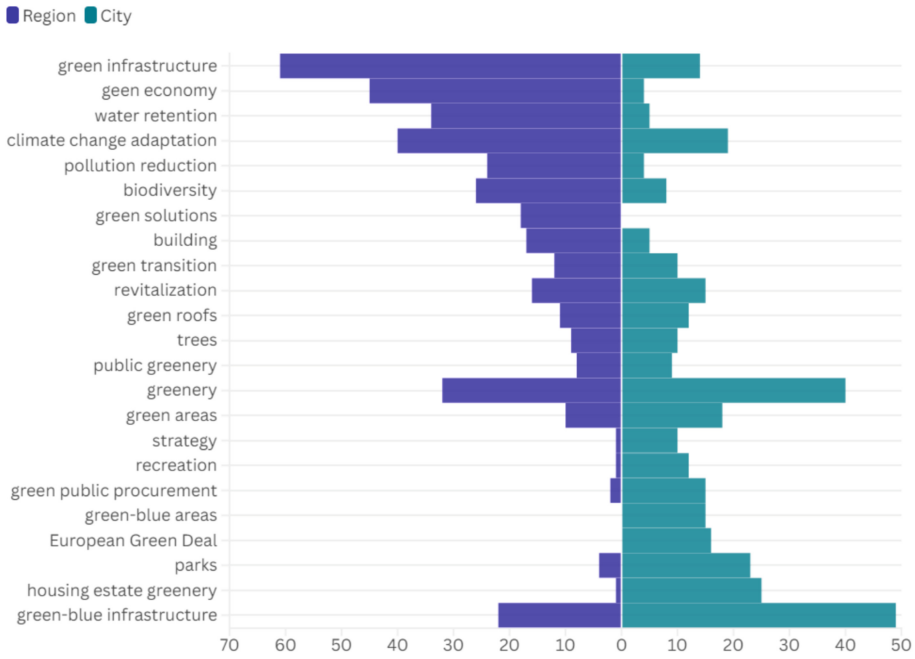


Fig. 3. Compliance and interconnectedness of regional and local strategies: Source: authors

Self-governing regions focus their attention mainly on those aspects of blue-green infrastructure which are focused on activities such as climate change adaptation, green economy, water retention, biodiversity, and pollution reduction. These terms can be perceived as fluid and developing based on the changing intensity of environmental challenges. On the other hand, cities usually connect their strategic goals with greenery, parks, and also infrastructure in general. The term greenery is also often used in regional documents, but the significant difference was identified through the words housing estate

greenery, public greenery and parks. However, a few similar practical priorities were identified, mainly green roofs, trees, green transition reflection and revitalisation. The focus of regions and cities is mainly on green aspects of blue-green infrastructure rather than blue or water elements of blue-green infrastructure.

5 Conclusion

Blue-green infrastructure is an important aspect of sustainable development. It has the potential to solve several issues and ensure benefits. Systematic planning and purposeful coordination are essential for the generation of benefits from blue-green infrastructure. The creation of policies and strategies to organise blue-green infrastructure can be seen as a key part of sustainable urban development.

The preliminary research results revealed the importance of blue-green infrastructure in the selected sample of local and regional self-government units. They also revealed the contexts in which the examined term is mostly mentioned. Based on the literature review and content analysis of local and regional strategic documents in selected self-government units, seven complementary components of the blue-green infrastructure have been identified: green infrastructure, administrative roles, green-blue infrastructure services, green-blue infrastructure benefits, climate change adaptation, community impact, and blue infrastructure.

The study's main benefit is its contribution to the important discussion about the sophistication of blue-green infrastructure as part of the resilient urban development and sustainable growth of local and regional communities. Although the study had many benefits, there are still limitations that need to be addressed in future research. Firstly, our findings were based on the intentional selection of self-government units in a specific country. The extension of our research can be furthered from the following aspects, mainly the broadening of the empirical evidence and further analysis of different approaches to blue-green infrastructure across municipalities, cities, regions and governments. Widening the scope to include more countries and conducting international comparisons could be very useful and valuable in this context.

References

Allen, C.R., Cumming, G.S., Garmestani, A.S., Taylor, P.D., Walker, B.H.: Managing for resilience. Wildl. Biol. **17**(4), 337–349 (2011). https://doi.org/10.2981/10-084

Ambily, P., Chithra, N.R., Mohammed, F.C.: A framework for urban pluvial flood resilient spatial planning through blue-green infrastructure. Inter. J. Disaster Risk Reduct. **103**(2024), article ID 104342 (2024). https://doi.org/10.1016/j.ijdrr.2024.104342

Andersson, E., Borgström, S., Haase, D., Langemeyer, J.: Urban resilience thinking in practice: ensuring flows of benefit from green and blue infrastructure. Ecol. Soc. **26**(4), article id 39 (2021). https://doi.org/10.5751/ES-12691-260439

Andersson, E.: Functional landscapes in cities: a systems approach. Landscape Ecol. Eng. **14**, 193–199 (2018). https://doi.org/10.1007/s11355-017-0346-6

Artmann, M., Kohler, M., Meinel, G., Gan, J., Ioja, I.: How smart growth and green infrastructure can mutually support each other - a conceptual framework for compact and green cities. Ecol. Indicators **96**(2019), 10–22 (2019). https://doi.org/10.1016/j.ecolind.2017.07.001

ARUP.: Cities Alive: Rethinking green infrastructure. https://www.arup.com/globalassets/downloads/insights/cities-alive-rethinking-green-infrastructure.pdf. Accessed 07 Feb 2025

Azadgoleh, M.A., Taghavijeloudar, M., Mohammadi, M.M., Ahangar, A.K., Yaqoubnejad, P.: A novel blue-green infrastructure for providing potential drinking water source from urban stormwater through a sustainable physical-biological treatment. Water Res. **273**(2025), article ID 123029 (2025). https://doi.org/10.1016/j.watres.2024.123029

Bears, R. C.: From traditional grey infrastructure to blue-green infrastructure. In: Bears, R.C. (eds.) Blue and Green Cities, pp. 1–41. Palgrave Macmillan, London (2018) https://doi.org/10.1057/978-1-137-59258-3_1

Boogaard, F.: From grey to green infrastructure in a changing climate. In: Bremer, S., Wardekker, A. (eds.) Changing Seasonality: How Communities are Revising their Seasons, Berlin, Boston: De Gruyter, pp. 217–224 (2024). https://doi.org/10.1515/9783111245591

Bray, I., Sinnett, D., West, H.: Exploring green, blue, grey infrastructure and neighbourhood satisfaction, quality of life, health and wellbeing in Bristol, UK. In: GAME CHANGER? Planning for just and sustainable urban regions - book of abstracts

Breuste, J., Artmann, M., Li, J., Xie, M.: Special issue on green infrastructure for urban sustainability. J. Urban Planning Develop. **141**(3) (2015). https://doi.org/10.1061/(ASCE)UP.1943-5444.000029

Bugnot, A.B., Mayer-Pinto, M., Johnston, E.L., Schaefer, N., Dafforn, K.A.: Learning from nature to enhance Blue engineering of marine infrastructure. Ecol. Eng. **120**(2018), 611–621 (2018). https://doi.org/10.1016/j.ecoleng.2018.03.012

Bystrica, B.: Information about the current activities of the city of Banská Bystrica in the area of the environment. https://cdn.banskabystrica.sk/2022/09/Informacia-o-aktualnych-aktivitach-mesta-Banska-Bystrica-v-oblasti-zivotneho-prostredia-31.8.2022-MsZ.pdf. Accessed 07 Feb 2024

Capobianco, V., et al.: The potential use of nature-based solutions as natural hazard mitigation measure for linear infrastructure in the Nordic Countries. Geoenviron Disasters **11**, article ID 27 (2024). https://doi.org/10.1186/s40677-024-00287-4

Chatzimentor, A., Apostolopoulou, E., Mazaris, A.D.: A review of green infrastructure research in Europe: challenges and opportunities. Landscape and Urban Planning **198**(2020), article ID 103775 (2020). https://doi.org/10.1016/j.landurbplan.2020.103775

Cheng, Y., et al.: Where and how to cool through blue infrastructure? large lake groups to ameliorate urban overheating in a typical inland multi-lake megacity. Sustainable Cities Soc. **98**(2023), article ID 104869 (2023). https://doi.org/10.1016/j.scs.2023.104869

Choi, Ch., Berry, P., Smith, A.: The climate benefits, co-benefits, and trade-offs of green infrastructure: a systematic literature review. J. Environ. Manag. **291**(2021), article ID 112583 (2021). https://doi.org/10.1016/j.jenvman.2021.112583

Cole, L.B., McPhearson, T., Herzog, C. P., Russ, S.: Green infrastructure. In: Russ, A., Krasny, M. E. (eds.) Urban Environmental Education Review, pp. 261–270. Cornell University Press, Ithaca, NY (2017). https://doi.org/10.1515/9781501712791-029

Cook, L.M., et al.: Towards the intentional multifunctionality of urban green infrastructure: a paradox of choice? NPJ Urban Sustainability. **4**, 12 (2024). https://doi.org/10.1038/s42949-024-00145-0

Das, M.: Green and blue infrastructure in high-density megacities of global south: a critical evaluation of benefits and obstacles for enhancing the well-being of Hong Kong's aging population. Developm. Environ. Sci. **15**, 575–594 (2024). https://doi.org/10.1016/B978-0-443-21948-1.00027-3

de Macedo, L.S.V., Picavet, M.E.B., de Oliveira, J.A.P., Shih, W.: Urban green and blue infrastructure: a critical analysis of research on developing countries. J. Cleaner Product. **313**(2021), article ID 127898 (2021). https://doi.org/10.1016/j.jclepro.2021.127898

de Oliveira, J.A.P., Bellezoni, R.A., Shih, W., Bayulken, B.: Innovations in urban green and blue infrastructure: tackling local and global challenges in cities. J. Cleaner Product. **362**(2022), article ID 132355 (2022). https://doi.org/10.1016/j.jclepro.2022.132355

de Rijke, C.A., Lim, N.J., Iqbal, A., Brandt, A.A., Sahlin, E.A.U.: A systematic review of blue-green infrastructure's role and relevance in the mitigation and management of climate-induced hazards in x-minute cities. Planning Pract. Res., 1–29 (2025). https://doi.org/10.1080/02697459.2025.2516529

Deak, J., Bucht, E.: Planning for climate change: the role of indigenous blue infrastructure, with a case study in Sweden. Town Planning Rev. **82**(6), 669–685 (2011). https://doi.org/10.3828/tpr.2011.38

European Commission.: Green and Blue Infrastructures (2021). https://knowledge4policy.ec.europa.eu/glossary-item/green-blue-infrastructures_en. Accessed 07 Feb 2025

European Commission.: Green infrastructure (2025). https://environment.ec.europa.eu/topics/nature-and-biodiversity/green-infrastructure_en. Accessed 05 Feb 2025

European Commission.: Nature-based solutions (2025). https://research-and-innovation.ec.europa.eu/research-area/environment/nature-based-solutions_en. Accessed 07 Jan 2025

European Union.: Green and blue infrastructure. https://www.interregeurope.eu/find-policy-solutions/policy-briefs/green-and-blue-infrastructure. Accessed 07 Feb 2025

Fenoglio, M.S., González, E., Calviño, A., Videla, M.: Urban insect communities in the neotropics: a systematic literature review and a green path to promote biodiversity conservation. In: León-Cortés, J.L., Córdoba-Aguilar, A. (eds.) Insect Decline and Conservation in the Neotropics, pp. 163–188. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-49255-6_8

Ferreira, J. C., dos Santos, D. C., Campos, L. C.: Blue-green infrastructure in view of integrated urban water management: a novel assessment of an effectiveness index. Water Res. **257**(2024), article ID 121658 (2024). https://doi.org/10.1016/j.watres.2024.121658

Garmendia, E., Apostolopoulou, E., Adams, W.M., Bormpoudakis, D.: Biodiversity and Green Infrastructure in Europe: boundary object or ecological trap? Land Use Policy **56**(2016), 315–319 (2016). https://doi.org/10.1016/j.landusepol.2016.04.003

Grabowski, Z.J., McPhearson, T., Marissa Matsler, A., Groffman, P., Pickett, S.T.A.: What is green infrastructure? a study of definitions in US city planning. Front. Ecol. Environ. **20**(3), 152–160 (2022). https://doi.org/10.1002/fee.2445

GreenBlue Urban.: Why Green and Blue? https://greenblue.com/gb/about-us/why-green-and-blue/. Accessed 07 Feb 2025

Haase, D.: Reflections about blue ecosystem services in cities. Sustainability Water Quality Ecol. **5**(2015), 77–83 (2015). https://doi.org/10.1016/j.swaqe.2015.02.003

Hanna, E., Bruno, D., Comín, F.A.: The ecosystem services supplied by urban green infrastructure depend on their naturalness, functionality and imperviousness. Urban Ecosystems. **27**, 187–202 (2024). https://doi.org/10.1007/s11252-023-01442-9

Hislop, M., Scott, A.J., Corbett, A.: What does good green infrastructure planning policy look like? developing and testing a policy assessment tool within Central Scotland UK. Plan. Theory Pract. **20**(5), 633–655 (2019). https://doi.org/10.1080/14649357.2019.1678667

Iojă, C. I., Badiu, D.L., Haase, D., Hossu, A. C., Niță, M. R.: How about water? Urban blue infrastructure management in Romania. Cities. **110**(2021), article ID 103084 (2021). https://doi.org/10.1016/j.cities.2020.103084

Johns, C.M.: Understanding barriers to green infrastructure policy and stormwater management in the City of Toronto: a shift from grey to green or policy layering and conversion? J. Environ. Planning Manage. **62**(8), 1377–1401 (2019). https://doi.org/10.1080/09640568.2018.1496072

Jones, L., Fletcher, D., Fitch, A., Kuyer, J. Dickie, I.: Economic value of the hot-day cooling provided by urban green and blue space. Urban Forestry & Urban Greening **93**(2024), article ID 128212 (2024). https://doi.org/10.1016/j.ufug.2024.128212

Kernaghan, S., Sturgeon, A.: Valuing nature as urban health infrastructure: a pathway to a nature positive city. Oxford Open Infrastruct. Health. **2**(2024), article ID ouae007 (2024). https://doi.org/10.1093/ooih/ouae007

Kim, J., Ewing, R., Rigolon, A.: Does green infrastructure affect housing prices via extreme heat and air pollution mitigation? a focus on green and climate gentrification in Los Angeles County, 2000–2021. Sustainable Cities Soc. **102**(2024), article ID 105225 (2024). https://doi.org/10.1016/j.scs.2024.105225

Korkou, M., Tarigan, A. K. M., Hanslin, H. M.: The multifunctionality concept in urban green infrastructure planning: a systematic literature review. Urban Forestry & Urban Greening **85**(2023), article ID 127975 (2023). https://doi.org/10.1016/j.ufug.2023.127975

Kronenberg, J., et al.: The thorny path toward greening: unintended consequences, tradeoffs, and constraints in green and blue infrastructure planning, implementation, and management. Ecol. Soc. **26**(2), 36 (2021). https://doi.org/10.5751/ES-12445-260236

Langemeyer, J., Baró, F.: Nature-based solutions as nodes of green-blue infrastructure networks: a cross-scale, co-creation approach. Nat.-Based Solut. **1**(2021), article ID 100006 (2021). https://doi.org/10.1016/j.nbsj.2021.100006

Lennon, M.: Green infrastructure and planning policy: a critical assessment. Local Environ. **20**(8), 957–980 (2014). https://doi.org/10.1080/13549839.2014.880411

Lin, Y., Wang, Z., Jim, C. Ch., Li, J., Deng, J., Liu, J.: Water as an urban heat sink: blue infrastructure alleviates urban heat island effect in mega-city agglomeration. J. Cleaner Product. **262**(2020), article ID 121411 (2020). https://doi.org/10.1016/j.jclepro.2020.121411

Liquete, C., et al.: Mapping green infrastructure based on ecosystem services and ecological networks: a Pan-European case study. Environ Sci Policy **54**(2015), 268–280 (2015). https://doi.org/10.1016/j.envsci.2015.07.009

Lõhmus, M., Balbus, J.: Making green infrastructure healthier infrastructure. Infect. Ecol. Epidemiológa **5**(1), article ID 30082 (2015). https://doi.org/10.3402/iee.v5.30082

Macrae, C.: Moments of resilience: time, space and the organisation of safety in complex sociotechnical systems. In: Wiig, S., Fahlbruch, B. (eds.) Exploring Resilience: A Scientific Journey from Practice to Theory, pp. 15–23. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-03189-3_3

Marot, N., Kostanjšek, B., Penko Seidl, N., Harfst, J.: Implementing the green infrastructure concept in practice: an example of the EU-strategy for Alpine Region. J. Environ. Planning Policy Manage. **26**(4), 325–338 (2024). https://doi.org/10.1080/1523908X.2024.2383685

Matsler, A.M., Meerow, S., Mell, I.C., Pavao-Zuckerman, M.A.: A 'green' chameleon: exploring the many disciplinary definitions, goals, and forms of green infrastructure. Landscape Urban Planning **214**(2021), article ID 104145 (2021). https://doi.org/10.1016/j.landurbplan.2021.104145

Meng, F., et al.: Quantification of the food-water-energy nexus in urban green and blue infrastructure: a synthesis of the literature. Resources, Conservat. Recycling **188**(2023), article ID 106658 (2023). https://doi.org/10.1016/j.resconrec.2022.106658

Molné, F., Donati, G.F.A., Bollinger, J., Fischer, M., Maurer, M., Bach, P. M.: Supporting the planning of urban blue-green infrastructure for biodiversity: a multi-scale prioritisation framework. J. Environ. Manag. **342** (2023), Article ID 118069 (2023). https://doi.org/10.1016/j.jenvman.2023.118069

Monaco, S.: SDG 9. Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation. In: Monaco, S. (eds.) Identity, Territories, and Sustainability: Challenges and Opportunities for Achieving the UN Sustainable Development Goals, pp. 87–93. Emerald Publishing Limited, Leeds (2024). https://doi.org/10.1108/978-1-83797-549-520241010

OECD.: Promoting nature-based solutions in municipalities in Hungary (2023). https://www.oecd.org/en/publications/promoting-nature-based-solutions-in-municipalities-in-hungary_d81fb09f-en.html. Accessed 07 Jan 2025

OECD.:Environmentally sustainable and climate-resilient infrastructure (2022). https://infrastructure-toolkit.oecd.org/infragov-theme/green/. Accessed 07 Feb 2025

Ojo, B.: Strategies for the optimization of critical infrastructure projects to enhance urban resilience to climate change. J. Sci. Eng. Res. **11**(6), 107–123 (2024)

Pauleit, S., et al.: Urban landscapes and green infrastructure. Oxford Res. Encyclopedia Environ. Sci. (2017). https://doi.org/10.1093/acrefore/9780199389414.013.23

Pereira, C., Flores-Colen, I., Mendes, M.P.: Guidelines to reduce the effects of urban heat in a changing climate: green infrastructures and design measures. Sustain. Dev. **32**(1), 57–83 (2023). https://doi.org/10.1002/sd.2646

Perrelet, K., Moretti, M., Dietzel, A., Altermatt, F., Cook, L. M.: Engineering blue-green infrastructure for and with biodiversity in cities. NPJ Urban Sustainability **4**(27), article ID 27 (2024). https://doi.org/10.1038/s42949-024-00163-y

Pinto, L. V., Inácio, M., Pereira, P.: Green and blue infrastructure (GBI) and urban nature-based solutions (NbS) contribution to human and ecological well-being and health. Oxford Open Infrastructure Health **1**(2023), ouad004 (2023). https://doi.org/10.1093/ooih/ouad004

Romańczak, A., Bratkowski, J., Sadowski, M.: Accessibility of green infrastructure at the city scale on the example of Łomża and Siedlce. Environ. Protect. Nat. Resour. **35**(1), 1–16 (2024). https://doi.org/10.2478/oszn-2024-0001

Rouse, D.C., Bunster-Ossa, I.: Green infrastructure: a landscape approach. In: Hamin Infield, E.M., Abunnasr, Y., Ryan, R.L. (eds.) Planning for Climate Change: A Reader in Green Infrastructure and Sustainable Design for Resilient Cities. Routledge, New York (2018). https://doi.org/10.4324/9781351201117

Scott, A., Hislop, M.: What does good green and blue infrastructure policy look like: a comparative assessment of UK national planning guidance. Urban Forestry & Urban Greening **99**(2024), article ID 128440 (2024). https://doi.org/10.1016/j.ufug.2024.128440

Severin, A., Michaliková, M.: Green and blue infrastructure: A Policy Brief from the Policy Learning Platform for a greener Europe. https://www.interregeurope.eu/sites/default/files/2024-09/Policy%20brief%20on%20Green%20and%20blue%20infrastructure.pdf. Accessed 10 Jan 2025

Shi, L.: Beyond flood risk reduction: How can green infrastructure advance both social justice and regional impact? Socio-Ecol. Pract. Res. **2**, 311–320 (2020). https://doi.org/10.1007/s42532-020-00065-0

Simon, D., et al.: Urban governance of and for Urban green and blue infrastructure. In: Shackleton, C.M., Cilliers, S.S., Davoren, E., du Toit, M.J. (eds.) Urban Ecology in the Global South. Cities and Nature, pp. 403–431. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67650-6_16

Sun, L., Xia, J., She, D.: Integrating model predictive control with stormwater system design: a cost-effective method of urban flood risk mitigation during heavy rainfall. Water Resources Res. **60**(4), article ID e2023WR036495 (2024). https://doi.org/10.1029/2023WR036495

Tate, C., et al.: The contribution of urban green and blue spaces to the United Nation's Sustainable Development Goals: An evidence gap map **145**(2024), article ID 104706 (2024). https://doi.org/10.1016/j.cities.2023.104706

Tetritsky. P. E.: Ecological infrastructure and its role in sustainable urban development: analysis and perspectives. Visnyk of V. N. Karazin Kharkiv National University. Series Ecology. 30, 63–75. https://doi.org/10.26565/1992-4259-2024-30-05

van Renterghem, T., Botteldooren, D., Verheyen, K.: Road traffic noise shielding by vegetation belts of limited depth. J. Sound Vib. **331**(10), 2404–2425 (2012). https://doi.org/10.1016/j.jsv.2012.01.006

Veerkamp, C.J., Schipper, A.M., Hedlund, K., Lazarova, T., Nordin, A., Hanson, H.I.: A review of studies assessing ecosystem services provided by urban green and blue infrastructure. Ecosyst. Serv. **52**(2021), article ID 101367 (2021). https://doi.org/10.1016/j.ecoser.2021.101367

Vitulano, V.: Integrating green infrastructure in Italian urban plans. Lessons from Turin and Bologna. In: Urban Des. Planning **177**(2), 45–56 (2024). https://doi.org/10.1680/jurdp.22.00049

Vogel, M. J.: The Role of green roofs in mitigating urban heat islands in Berlin, Germany. J. Res. Soc. Sci. Humanit. **3**(11), 48–54 (2024). https://doi.org/10.56397/JRSSH.2024.11.06

Wang, J., Banzhaf, E.: Towards a better understanding of Green Infrastructure: a critical review. Ecol. Indicators **85**(2018), 758–772 (2018). https://doi.org/10.1016/j.ecolind.2017.09.018

World Bank Group.: What You Need to Know About Nature-Based Solutions to Climate Change (2022). https://www.worldbank.org/en/news/feature/2022/05/19/what-you-need-to-know-about-nature-based-solutions-to-climate-change. Accessed 07 Jan 2025

Ying, J., Zhang, X., Zhang, Y., Bilan, S.: Green infrastructure: systematic literature review. Econ. Res. Ekonomska Istraživanja **35**(1), 343–366 (2022). https://doi.org/10.1080/1331677X.2021.1893202

The Challenges of Urban Regeneration Between Conservation and Transformation in Coastal Areas. The Case Study of La Caletta, Sardinia, Italy

Ginevra Balletto(✉) and Mara Ladu

DICAAR - Department of Civil, Environmental Engineering and Architecture, University of Cagliari, Cagliari, Italy
{balletto,mara.ladu}@unica.it

Abstract. Balancing the conservation of modern architecture's legacy with the demands of contemporary urban transformation represents a key challenge for urban regeneration, particularly in coastal areas affected by tourism development pressure. The coastal settlement of La Caletta, in the municipality of Siniscola, on the central-eastern coast of Sardinia (Italy), represents one of the most interesting contexts for observing, on a local scale, the interplay between the major ongoing transitions and the specificities of Sardinian coastal landscapes. In this context, the history of tourism development, the expansive urban growth and the consequent transformation, as well as demographic changes have shaped a complex geography, in which the relationship between the main urban center (Siniscola) and the coastal settlement (La Caletta) has progressively been redefined. This paper investigates the opportunities offered by sustainable urban regeneration, understood as a systematic set of actions aimed at qualitative renewal and strengthening of both material and immaterial infrastructures (physical space and digital environment) in order to improve the quality of life of residents and city users. The study takes as its starting point the experience of the Villaggio Milano, designed by Carlo De Carli in the late 1940s, that can be considered a significant combination between residential and tourist demand, modern architecture, characteristics of the coastal landscape, and connections with the broader geographical context.

Keywords: Urban Regeneration · Coastal areas · Sustainable Spatial planning

1 Introduction

Urban regeneration has been at the core of disciplinary debate for more than twenty years, representing one possible response to the environmental, economic, and social challenges facing the contemporary city [1–3].

In Italy, until the 1990s, spatial planning was dominated by the management of the building expansion. With the shift from an urbanism of urban growth to one of demographic decline and urban regeneration, new priorities came to dominate the spatial

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 610–621, 2027.
https://doi.org/10.1007/978-3-032-30521-3_38

planning domine: the need to reuse the existing city to reduce land consumption, mitigate climate change, and improve urban resilience; the importance of moving from recovery or redevelopment actions, often limited to the physical-building dimension, to the concept of regeneration, based on a systemic approach integrating social, economic, and environmental policies; the urgency of rethinking other parts of the city beyond the historic center, such as peripheral areas, urban "voids," and brownfield sites (former industrial areas, railway yards, barracks), with particular attention to the public city [4]; the awareness of the role of participation and governance in managing complex processes involving a plurality of stakeholders (public, private, and third sector), together with the updating of the regulatory framework and spatial planning tools to make them consistent with the principles of sustainability [5, 6].

The debate prominently involves major sector organizations, including the National Institute of Urban Planning (INU), the Italian Society of Urban Planners (SIU), the Association of Disused Urban Areas (AUDIS), as well as the National Council of Engineers (CNI) and the National Council of Architects, Planners, Landscape Architects and Conservators (CNAPPC).

INU's commitment to urban governance (at local, regional and national scale) has promoted an integrated approach to the design of the existing city, synthesized in a series of official documents that highlight how urban regeneration should be understood as a new frontier of Italian urbanism for acting on the social and environmental fabric, restoring dignity to peripheral and degraded areas, and thereby countering land consumption [7]. The ten proposals presented by INU in 2017 outline an operational strategy to turn urban regeneration from an intention into a structural practice [8]. The main fields of action concern: operational tools (the establishment of a regular state fund to finance integrated programs addressing social distress and seismic/hydraulic risk); necessary reforms in housing policy and urban standards (to allow a shift from the concept of "quantity" to the "quality" of contemporary spaces and services [9], including smart grids, green management, and climate resilience); land consumption (downsizing municipal local plans and updating outdated zoning previsions); and experimentation and system action to coordinate national policies on building rehabilitation and urban regeneration under a single coordinating authority.

Many of these principles are also expressed in the document of the CNAPPC, which proposed a National Plan for Sustainable Urban Regeneration (2012) [10] as a priority strategy to overcome the construction crisis and relaunch the Italian economy. Here, the key proposals concern: the renewed role of demolition-and-reconstruction projects, using urban equalization and compensation tools; the enhancement of public real-estate assets; the need to promote ecological transition through the creation of eco-neighborhoods ensuring water and energy efficiency and seismic safety; a general renewal and strengthening of urban welfare through the upgrading of public spaces and digital infrastructure to foster social cohesion and inclusion; the updating of the regulatory framework through simplified procedures, tax incentives, and coordinated national governance.

These issues are also reaffirmed in the AUDIS Charter for Urban Regeneration (2015) [11], which helps consolidate a framework of principles supporting city makers. (public and private actors).

More recently, INU proposed guidelines for addressing the post-Covid-19 pandemic phase in an extremely dramatic moment for the country [12], recognized urban regeneration as a tool for creating healthy, accessible, and inclusive environments, in line with the concept of the 15-min city, where public spaces are redesigned to ensure local services, widespread greenery, and softer mobility, reducing spatial inequalities that the health crisis made even more evident.

At the XXIII SIU National Conference (2021), dedicated to the regeneration of urban space and social transformation, urban development schemes characterized mostly by episodic interventions were strongly discussed for their limited ability to fit into a broad strategic vision, which leads to the exclusion of certain parts of the city and, consequently, social marginalization. Reflections thus return to the issues of urban marginality and housing, the relationship between the urban plan and the urban project, and the operational methods that make it possible to engage with territory through policies and practices based on the broadest participation [13].

INU's most recent 2025 motion [14] reproposes urban regeneration as an opportunity to address the challenges of sustainable urban planning through an integrated approach, recognizing the centrality of the planning tool, the local plan, in identifying the areas to be subjected to regeneration interventions, in defining the necessary incentives, and in enabling the transfer of building volumes and development rights under public leadership. It also calls for a national law guaranteeing continuous financial resources, tax relief, and a comprehensive reform of urban standards.

In this regard, the INU's legislative commitment is also evident in its various proposals for a new Urban Planning Law that would provide the Italian regions with a single regulatory instrument capable of renewing and coordinating regional regulatory frameworks on territorial governance, in line with the sustainable development goals.

The more recent bill on urban regeneration discussed by the Italian government (DDL Rigenerazione urbana 2025) is largely the outcome of the disciplinary advances described above. The text introduces a national strategy to bring net land consumption to zero by 2050 and establishes a national fund to finance rehabilitation of the built environment, energy efficiency upgrades, and seismic safety measures in municipalities. It also introduces volumetric bonuses for demolition and reconstruction, as well as specific incentives for relocating buildings from areas exposed to hydrogeological risk.

Significant procedural simplifications and tax advantages are also provided, with the primary aim of improving quality of life in urban environments by integrating environmental sustainability, social housing, and soft mobility.

Within this framework, this contribution explores the opportunities offered by urban regeneration, taking as its starting point the case of Villaggio Milano in La Caletta di Siniscola (Sardinia, Italy), one of the first tourist accommodation village developed in Sardinia and designed in the late 1940s by Carlo De Carli. Although it has undergone significant transformations over time, which have strongly damaged its memory, this valuable project can be considered a significant combination between residential and tourist demand, modern architecture, characteristics of the coastal landscape, and connections with the broader geographical context. In this sense, La Caletta is today one of the most interesting contexts for observing, at a local scale, the interplay between the major transitions underway and the specific features of Sardinian coastal landscapes.

In this context, the history of tourism development, the expansive urban growth and the consequent transformation, as well as demographic changes have shaped a complex geography, in which the relationship between the main urban center (Siniscola) and the coastal settlement (La Caletta) has progressively been redefined.

The manuscript is structured in 6 sections: the introduction is followed by an overview of the legacy of theVillaggio Milano (Sect. 2), between memory and transformation (Sect. 3), the challenges of urban regeneration (Sect. 4) and the contextual value of the symbolic and cultural dimension (Sect. 5) in guiding future urban renewal policies. The final section is dedicated to the conclusion of the study (Sect. 6).

2 Villaggio Milano. The Legacy of the Past

From its origins, La Caletta has occupied a threshold position between the Siniscola hinterland and the sea, functioning at once as a landing place, a seaside resort, and, increasingly, a place of permanent residence (Fig. 1).

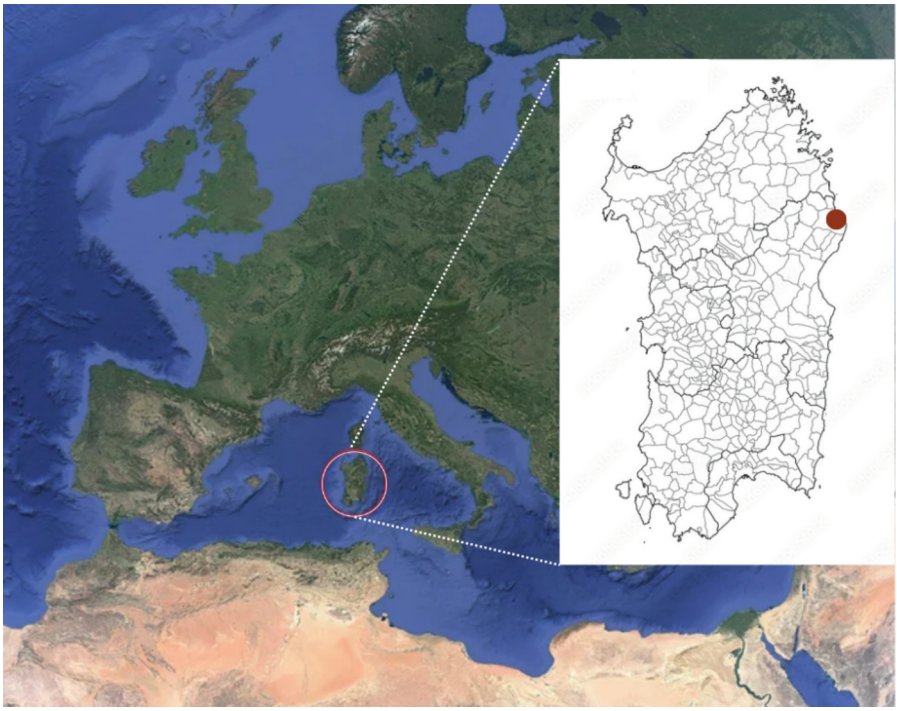


Fig. 1. The study area. La Caletta di Siniscola in Sardinia, Italy (in red)

Its relationship with the municipal center has historically been marked by a clear functional distinction: the inland town as the seat of services and historical identity, and the coastal district as a space for leisure and seasonality. In the last twenty years, the port of La Caletta has mainly served as a fishing port and a tourist port (Fig. 2).

In recent decades, however, demographic change and patterns of land use have challenged this dichotomy, giving rise to a network of mobility and permanence that connects the two urban centers daily. La Caletta is no longer merely a "holiday place," but a fragment of coastal urbanity in which residents, commuters, tourists, and new temporary inhabitants interact.



Fig. 2. The tourist harbor of La Caletta, Siniscola (Sardinia, Italy)

Within this framework, the tourist accommodation village (Villaggio Milano) designed in the late 1940s by Carlo De Carli, emerges as a foundational episode in Sardinian tourist modernity. It was the island's first tourist settlement, conceived as a device capable of combining architectural quality, attention to the landscape, and experimentation with forms of temporary dwelling. The layout of the village reflects a particular sensitivity to the site's morphology: buildings are organized in relation to the topography, open spaces are conceived as extensions of the housing units, and pedestrian paths create a sequence of views, thresholds, and vistas toward the sea. In this way, the project departs both from the logic of the continuous built front along the coast and from the random dispersion of isolated volumes, proposing a balance between density and permeability. It embodies an idea of vacation that engages with the landscape not in terms of mere occupation of space, but as a dialogue with nature, topography, and light. The composition of the buildings, the use of materials, and the arrangement of open spaces and pathways create a settlement system that seeks to enhance the threshold

condition between the built environment and the sea, between the individual experience of temporary residence and the collective experience of shared spaces.

However, the legacy of Villaggio Milano has not remained isolated within the landscape of La Caletta. It can be interpreted as a forward-looking laboratory for many of the issues now central to urban regeneration: sustainable density, the relationship between private and shared spaces, and the ability of existing architectural assets to activate new forms of use and place-based interpretation. Its position within the Siniscola coastal landscape and its formal recognizability make it a potential anchor for enhancement policies that do not stop at passive preservation but instead promote forms of reuse compatible with contemporary needs. This, among other things, implies reflection on the role of architectural memory and on the possibility of building interpretive pathways for modern architectures, involving residents, visitors, and professional communities.

3 Between Memory and Transformation

From the 1960s onward, and even more intensely during the 1970s and 1980s, the coastal strip of La Caletta experienced a progressive spread of settlement, driven by waves of demand for second homes and by the growing centrality of seaside tourism in the local economy. The result is a layered urban fabric in which high-quality design interventions coexist with building additions developed mainly according to speculative or purely quantitative logics. Diachronic analyses of the different evolutionary phases—from the 1950s to the 1970s, and then through more recent developments between 2017 and 2024—highlight the shift from a relatively compact settlement model to fragmented forms of growth, often poorly integrated with one another.

Reading these changes through the digital land cover platform, obtained by comparing historical image series and territorial data, makes it clear how urban expansion has affected the original environmental mosaic. Natural and semi-natural surfaces have been progressively replaced by low-density residential fabrics, local roads, parking areas, and hospitality facilities, with a consequent fragmentation of open spaces and ecological corridors.

This process is common to many Mediterranean coastlines, but in La Caletta it takes on particular significance because it concerns a context in which landscape quality constitutes the main resource on which tourist attractiveness has been built.

The central issue for regeneration is not so much "how to grow further," but rather how to reconnect, selectively densify, and restore continuity and meaning to the spaces that are now interstitial. The notion of "dispersion and densification" can be understood not only as a description of a past process, but also as an interpretive approach for future strategies. In recent years, dispersion has meant settlement sprawl and land consumption; densification can today mean, by contrast, a reasoned concentration of functions and interventions within already urbanized fabrics, aimed at improving accessibility, public space quality, and mixed use, in line with the paradigms of the sustainable city [15, 16].

4 The Challenges of Urban Regeneration

The regeneration of the La Caletta tourist settlement depends on the development of a systemic project for open space, understood as a primary infrastructure that permeates and reshapes the consolidated urban fabric and its future expansions [17].

Conceiving open space as a connective tissue and a proximity hub integrating different functions (sports, urban gardens, sustainable mobility stations, and outdoor coworking spaces) means recognizing the value of the elements of the urban fabric in all their forms [18].

According to the definition proposed by the Public Space Charter (2013) [19] later by the Global Public Space Toolkit (2016) [20]: "Public space is any place that is publicly owned or publicly used, accessible and usable by all free of charge or without profit-making purposes. […] a key element of individual and social well-being […]. Public spaces consist of open spaces (such as streets, sidewalks, squares, gardens, parks) and covered spaces created without profit-making purposes for the benefit of all (such as libraries, museums)." However, for public spaces to effectively constitute "the physical network and support for the movement and stopping of people and vehicles, on which the functioning and vitality of cities depend," it is necessary to ensure that all their elements form a continuous, articulated, and integrated system [21].

In this sense, the recovery of urban voids, the reorganization of pedestrian and cycle routes, and the strengthening of the relational nodes between La Caletta and the agricultural and mountainous hinterland of Siniscola become strategic actions (Fig. 3).

Systematically coordinating these interventions can reduce pressure on further undeveloped areas while helping to recombine a fragmented urban landscape, creating an infrastructure capable of responding simultaneously to the challenges of contemporary life: adaptation to climate change, social inclusion, and economic regeneration.

Among the most recent approaches, Nature-Based Solutions (NbS) [22] should be mentioned: intelligent design solutions to support sustainable water management, including reducing the load on sewage networks (rain gardens), and reducing urban heat islands (urban forestry), with particular attention to biodiversity conservation.

From a social perspective, the open-space project aims to restore meaning to places and create new spaces that enhance memory and local community relations, while respecting the principle of "Accessibility for all" [23], namely the design of inclusive spaces accessible to all age groups, avoiding gentrification processes.

From an economic perspective, significant resources could also come from the proceeds of the tourist tax which, according to the legal framework (Art. 4 of Legislative Decree No. 23/2011), must be used to support the governance of tourist flows in high-density destinations, thereby financing interventions for the maintenance, use, and recovery of local cultural and environmental assets—including sustainable waste management—urban renewal and public realm improvement, the enhancement and strengthening of local public services (e.g., mobility and transport) and cultural events, to the benefit of both the resident community and city users [24].

A tourist tax can help curb overtourism easing pressure toward gentrification only if it is carefully structured, fenced, and managed transparently with input from local communities. More precisely, main aspects concern finance resident-oriented services, combined with strong housing and land-use policies that explicitly prevent gentrification.

Revenues should mainly support residents and common-good infrastructure, such as relieving public transport (buses, metro, bike lanes), waste collection, public toilets, and park maintenance, as well as social housing schemes, rent-control-related programs, or rehabilitation of historic housing to expand affordable-housing stock to protect low- and middle-income residents or support long-standing tenants against rent hikes.

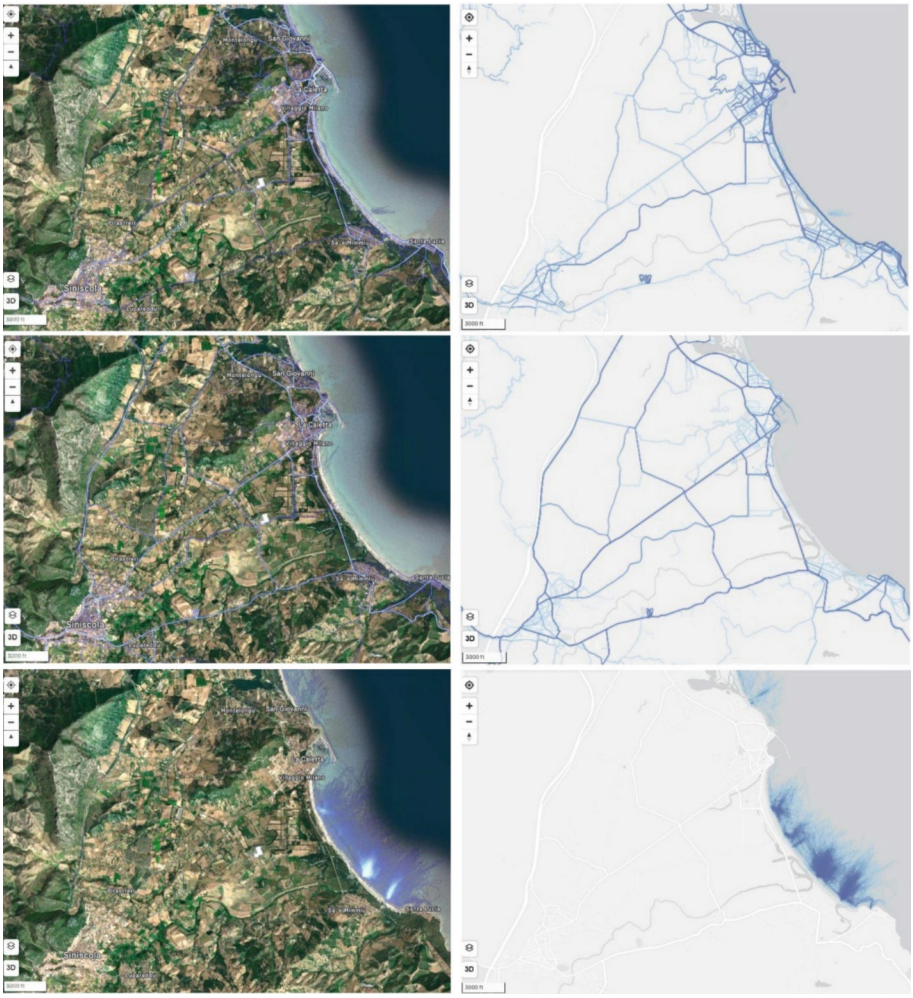


Fig. 3. The STRAVA heatmaps related to the main connections between La Caletta and Siniscola. The maps show the places where users record their sports activities: walking a), cycling b), and water sports c). Source: STRAVA.

An innovative element of the current framework is the digital transition, which introduces a new immaterial infrastructure [25]. Connectivity maps—showing fixed-line speeds and wi-fi coverage—reveal that accessibility is no longer only a matter of road

infrastructure, ports, and parking areas, but also of data networks and online services [26]. The availability of stable and fast connections directly affects the ability to attract new residents, remote workers, and companies linked to the knowledge economy. In a context such as La Caletta, digital transition could shape interesting future scenarios for the de-seasonalization of presence and for the integration of traditional tourism with new forms of staying that combines work and leisure. Digitalization, however, raises important questions about social equality. Differences in coverage and service quality across different parts of the municipality may reinforce existing inequalities, further marginalizing less well-served areas. For digital transition to become a lever for urban regeneration, it must be integrated into planning policies and local programming tools, making connectivity a structural dimension of territorial design.

5 Symbolic and Cultural Dimension

The symbolic and cultural dimension that helps define La Caletta's identity cannot be overlooked. The reference to André Pieyre de Mandiargues' novel The Sea Lily, set precisely in this coastal location, highlights how the place was already, by the mid-20th century, the setting for narratives intertwining nature, memory, and identity. The description of seascapes, atmospheres, and holiday experiences conveys an image of La Caletta that goes beyond its physical reality, projecting it into the collective imagination as a space of passage and transformation. This literary component, alongside the architectural history of Villaggio Milano and the evidence of industrial and artistic heritage (such as the Monteponi mining village and the frescoes by Aligi Sassu), helps build a symbolic capital on which regeneration policies can and should rely.

Regenerating La Caletta, in this sense, does not mean only restructuring buildings, improving infrastructure, and upgrading public spaces, but also reactivating narratives and reconnecting the community with the multiple legacies inhabiting the territory. The interplay between nature and artifice, between modern architecture and coastal landscape, and between memory and contemporary tourism offers the possibility of imagining enhancement paths that do not reduce the place to a mere tourist product [27, 28]. On the contrary, regeneration can become an opportunity to rethink the pact between residents, visitors, and the environment, restoring to the Siniscola coastline the role of a laboratory for new forms of coastal living, more attentive to the resources, fragilities, and opportunities of ongoing transitions.

6 Conclusions

The considerations developed in this study suggest the need to reorient territorial and urban planning instruments so that they treat regeneration as a strategic axis rather than a simple add-on chapter. First, the local plan can become the place where the value of Villaggio Milano and the historic coastal fabrics is formally recognized, with specific protection and reuse regulations that encourage recovery interventions, qualitative densification, and the enhancement of public spaces [29], while discouraging further land consumption. From this perspective, the regulation of consolidated areas and completion areas should be reconsidered considering the objectives of reducing settlement fragmentation, improving soft accessibility, and restoring ecological continuity.

Alongside structural planning, a crucial role can be played by integrated urban and territorial regeneration programs, developed through regional calls, European funds, and national initiatives dedicated to coastal areas and smaller urban settlements. Such programs could act on several fronts: public-space and waterfront redevelopment, recovery of underused buildings to be assigned to cultural, social, or diffused hospitality functions, energy-efficiency improvements, and seismic upgrading of the built heritage. Significant resources could also come from the proceeds of the tourist tax which in Italy has been constantly growing due to the increase in the number of tourists and the number of municipalities that introduced the tax. Despite the existence of legislation that limits the use of tourist tax funds to specific interventions in the tourism sector, a strategic vision should finance regeneration schemes focused on revitalizing local cultural and environmental assets.

Within this framework, policies for digital transition should be explicitly linked to regeneration objectives, making connectivity a lever for supporting new local economies, remote-working arrangements, and innovative services, while preventing coverage gaps from turning into further territorial inequalities. Finally, planning strategies should more decisively integrate the landscape and cultural dimension, enhancing the interplay between modern architecture, memory, natural heritage, and the literary imagination. This implies, on the one hand, close dialogue with landscape-planning instruments and conservation policies; on the other, the promotion of projects that connect places such as Villaggio Milano, the marina, coastal spaces, and inland routes, building thematic itineraries capable of strengthening La Caletta's identity [30]. In this sense, regeneration becomes an eminently political and cultural operation, requiring participatory processes, alliances among administrations, local communities, and economic actors, and the ability to use plans, programs, and policies in a coordinated way as instruments of a unique territorial vision.

Acknowledgments. This study is the result of the contribution to the conference "Villaggio Milano di Carlo De Carli. L'eredità del primo insediamento turistico della Sardegna per la rigenerazione del territorio e la valorizzazione del paesaggio a La Caletta di Siniscola" / "Villaggio Milano by Carlo De Carli. The legacy of Sardinia's first tourist settlement for the regeneration of the territory and the enhancement of the landscape in La Caletta di Siniscola" held on 20 September 2025 in La Caletta di Siniscola. The publication of the Conference Proceedings is funded by Fondazione di Sardegna, Annual Call 2026 – Local Development.

Author Contributions: G.B. and M.L wrote all sections. The two authors have read and agreed to the published version of the manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Said, Z.M., Dindar, S.: Key challenges and strategies in the evaluation of sustainable urban regeneration projects: insights from a systematic literature review. Sustainability **16**(22), 9903 (2024)

2. Errigo, M. F., Iacomoni, A.: Rigenerazione urbana e governo della città contemporanea. Le prospettive e le azioni in ambito ecologico-ambientale e urbanistico. I contributi. In Rigenerazione Urbana E Governo Della Città Contemporanea. La Prospettiva Ecologico-Ambientale. Piani E ProgettI, pp. 13–20 (2022).. Inu Edizioni
3. Losco, S., de Biase, C.: Consumo di suolo versus eco-planning. In: La declinazione della sostenibilità ambientale nella disciplina urbanistica, vol. 3, pp. 123–129. Planum Publisher e Società Italiana degli Urbanisti (2023)
4. Ladu, M.: Patrimonio immobiliare pubblico. Prospettive di rigenerazione per una città accessibile e di prossimità, vol. 2. UNICApress (2022).
5. Torelli, G.: La rigenerazione urbana nelle recenti leggi urbanistiche e del governo del territorio. Le Istituzioni Del Federalismo **3**, 651–679 (2017)
6. Di Lascio, F.: Governare la rigenerazione urbana. Rivista quadrimestrale di diritto dell'ambiente **2**(2022), 88–115 (2022)
7. Berretta, T., Desideri, F., Staltari, M.: Il progetto dello spazio pubblico, tra complessità e crisi ecologica: da sfida a opportunità per la rigenerazione urbana. Agathon: Inter. J. Archit. Art Design **16** (2024).
8. Istituto Nazionale di Urbanistica (INU), Rigenerazione urbana. Le 10 proposte dell'INU. 4 strumenti, 4 riforme, 1 sperimentazione, I azione di sistema (Novembre 2017): http://www.inu.it/wp-content/uploads/10_proposte_INU.pdf
9. Arcidiacono, A., Giaimo, C., Talia, M.: Priorità per il progetto urbanistico. Reti ambientali, nuovi standard e rigenerazione urbana. Accademia, 166–179 (2018).
10. CNAPPC Piano Nazionale per la Rigenerazione Urbana Sostenibile (2012). https://www.awn.it/attachments/article/731/CNAPPC_Piano_Nazionale_per_la_Rigenerazione_Urbana_Sostenibile.pdf
11. AUDIS La Carta AUDIS della Rigenerazione Urbana, Bologna (2015)
12. Istituto Nazionale di Urbanistica (INU) Le proposte dell'Istituto Nazionale di Urbanistica per il superamento dell'emergenza e il rilancio del Paese (2020). http://www.inu.it/wp-content/uploads/proposte-inu-15-maggio-2020.pdf
13. Caruso, N., et aI. Rigenerazione dello spazio urbano e trasformazione sociale. Introduzione. In: Caruso N., Pasqui G., Tedesco C., Vassallo I. (eds.) Rigenerazione dello spazio urbano e trasformazione sociale. Atti della XXIII Conferenza Nazionale SIU. DOWNSCALING, RIGHTSIZING. Contrazione demografica e riorganizzazione spaziale, Torino, 17–18 giugno 2021, vol. 05. Planum Publisher e Società Italiana degli Urbanisti, Roma-Milano, pp. 5–8 (2021)
14. Istituto Nazionale di Urbanistica (INU), Mozione sulla Rigenerazione urbana approvata all'unanimità. XXXII Congresso nazionale – Assemblea dei soci del 24 maggio (2025). http://www.inu.it/wp-content/uploads/allegato-4-inu-mozione-rigenerazione-urbana.pdf
15. United Nations. Transforming our world. The 2030 agenda for sustainable development. New York: Resolution adopted by the General Assembly on 25 September 2015 (2015). Retrieved 21 September 2018
16. Moreno, C. The 15-Minute city: a solution to saving our time and our planet. John Wiley & Sons (2024)
17. Cutini, V.: La rivincita dello spazio urbano: l'approccio configurazionale all'analisi e allo studio dei centri abitati (2010)
18. Marella, R.M.: Lo spazio urbano come bene comune. TECHNE–J. Technol. Archit. Environ. **28**, 28–34 (2024)
19. Carta dello Spazio Pubblico 2013, Biennale Spazio Pubblico Maggio, 2013, Biennale Spazio Pubblico» Carta dello Spazio Pubblico (2013)
20. United Nations Global Public Space Toolkit. From Global Principles to Local Policies and Practice. United Nations Human Settlements Programme UN-Habitat, SIDA, INU (2016)

21. Balletto, G., Sinatra, M., Ladu, M., Sechi, F., Borruso, G.: A methodological proposal to assess proximity in functional urban areas. In International Symposium: New Metropolitan Perspectives, pp. 70–79. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-74679-6_7
22. Ladu, M., Sulis, G.: Nature-Based Solutions per mitigare gli impatti della urbanizzazione: il caso del Parco di Molentargius-Saline. Reticula **29**, 44–59 (2022)
23. United Nations: Accessibility and Development: Mainstreaming Disability in the Post-2015 Development Agenda. Department of Economic and Social Affairs, New York (2015)
24. Balletto, G., Ladu, M., Battino, S., Attard, M.: A methodological perspective on understanding overtourism in Mediterranean Islands. the case of Sardinia region, Italy. City, Territory Archit. **12**(1), 31 (2025)
25. Borruso, G., et al.: Aree interne e sviluppo locale al tempo dei social network e media. Il caso del Sulcis–Iglesiente (Sardegna, Italia). Trieste, Italy: EUT Edizioni Università di Trieste (2022).
26. Ladu, M., Balletto, G., Paddeu, D., Corona, L.: Beyond spatial accessibility. redefining access to services through digital connectivity. Euro. Trans. **103**, 1–17 (2025)
27. Balletto, G., Borruso, G., Ladu, M., Milesi, A.: Smart and slow tourism. evaluation and challenges in Sardinia (Italy). In: International Conference on Innovation in Urban and Regional Planning, pp. 175–182. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-96985-1_20
28. Balletto, G., Ladu, M., Sinatra, M., Borruso, G.: Application of Geospatial analysis to tourist flows in the insular destinations. In: International Conference on Computational Science and Its Applications, pp. 241–252. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-65308-7_17
29. Ladu, M.: Rigenerare il patrimonio costruito, tra sviluppo turistica e domanda abitativa. In Santa Lucia. Paesaggi e passaggi. 3ª rassegna di studi storici e sociali in onore della dottoressa Antonietta Boninu, pp. 184–206. Editrice Taphros srls (2025)
30. Borruso, G., Balletto, G.: The port-city relations in the era of hybridization. In: International Conference on Computational Science and Its Applications, pp. 594–607. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-37123-3_41

Machine Learning Methods for Fire Risk Assessment in Wildland Urban Interfaces - Case Study Attica Region, Greece

Angeliki Papazoglou[1(✉)], Apostolos Lagarias[2], and Anastasia Stratigea[1]

[1] Department of Geography and Regional Planning, School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Athens, Greece
angeliki.papazoglou@hotmail.com, stratige@central.ntua.gr
[2] Department of Regional Planning and Development, School of Engineering, University of Thessaly, Volos, Greece
lagarias@uth.gr

Abstract. Wildfires' severity and frequency, further deteriorating due to Climate Change, are currently gaining great attention due to the devastating damages these may cause in the natural/built environment. Thus, proactive planning, targeting mitigation, preparedness, response and recovery, has become a necessity and needs to embrace Disaster Management perspectives, originating from wildfires. In support to this emergency, this study aims to develop a fire risk assessment framework relevant to Wildland-Urban Interfaces (WUIs), i.e., highly vulnerable areas due to their close proximity to forested ones. Towards this end, insights from the case study area – Attica region, Greece – and the inherent WUIs' characteristics has led to the establishment of an exhaustive fire-related variables' dataset. This has formed the ground for comparing robustness of a number of Machine Learning (ML) Methods towards fire risk assessment. Six literature-driven Individual ML Methods – Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Logistic Regression (LR), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) – and one Integrated Ensemble Method (Voting Ensemble) are used in this respect. Results obtained witnessed strong predictive performance of all methods, with XGBoost and LightGBM demonstrating the best overall results across multiple evaluation metrics. Worth noticing is also the minimal divergence between Test AUC and Spatial Cross-Validation AUC, indicating robustness to spatial generalization. Finally, contrasting fire risk assessment results to burned areas in the time span 2000–2024 revealed strong spatial alignment of all methods.

Keywords: Climate Change · Wildland Urban Interfaces (WUIs) · Fire Risk Assessment · Machine Learning Methods · Attica Region/Greece

1 Introduction

In the Climate Change (CC) era, incidents of natural disasters around the globe display escalating trends, raising thus concerns as to the severity, intensity and frequency of these phenomena; and the devastating consequences these cause in the natural/cultural

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 622–640, 2027.
https://doi.org/10.1007/978-3-032-30521-3_39

and built environment. Reports of IPCC on future climatic projections demonstrate the prevalence of hot extremes and extended heat waves, droughts and decreasing precipitation rates in the years to come, particularly in the Mediterranean, an area marked as a CC "hotspot" [1]. That said, weather conditions in particular are expected to amplify the number of fire incidents in the Mediterranean, thus extending the fire season and the number of consecutive days with high or extreme fire danger [2, 3], affecting region's vulnerability and exposure to forest fires.

The rising frequency and severity of natural disasters in the CC era has raised research interest and risk assessment endeavours in, among others, the Mediterranean Region [2]. Thus, apart from indulging into the specificities of Mediterranean as a CC hotspot, research interest and relevant works are also focusing on the human factor and, more precisely, on the continuously escalating, poorly-planned urban sprawl along the Mediterranean coast. This is due to urbanization trends and the rising pressure for expanding mass tourism accommodation infrastructure, with both reinforcing vulnerability and exposure to wildfire incidents [4]. Urban sprawl, in particular, has given rise to the *Wildland Urban Interfaces* (WUIs), lying at the heart of this work. These are land compartments in close proximity to fire prone forested and vegetated areas. In fact, WUIs in the Mediterranean have so far witnessed the devastating consequences of severe wildfire outbreaks in their natural and socioeconomic environment. For instance Greece, along with other Mediterranean countries, every fire season is confronted with extreme wildfire events and their repercussions in both forested areas (e.g., Evia Island in 2021, area of Alexandroupolis in 2023) [5, 6] and the neighbouring urban constellations. Attica region, i.e., the case study of this research, is also one of the most affected areas in this respect. This area is marked by expanded WUIs throughout its territory, which have gone through severe wildfire disasters, with immeasurable repercussions in both the natural and built environment, but also human losses and deterioration of the quality of life of local population [2, 6, 7].

The escalating number of disaster incidents in the CC era and their severe consequences highlight the need to assess risk and set up proactive strategies in pursuing resilience objectives of both the natural and the man-made environment. A great ally in this effort is the immense technological developments, which have shifted scientific interest towards state-of-the-art and more reliable methodologies in support of disaster management enhancement. Speaking of wildfires, the need for more efficient fire risk assessment methodologies that are capable of coping with complexity inherent in this research topic is particularly evident [8, 9]. The advent and rapid growth of Artificial Intelligence (AI) and Machine Learning (ML) tools represent a valuable step towards this direction, proved also by their widespread adoption across relevant scientific fields. Due to their capability to manage large volumes of complex data [10], minimize human bias and deliver promising and reliable results in complex research endeavours [9, 11], state-of-the-art methodologies are increasingly replacing conventional ones. Enabled by technological evolution and computing power, the adoption of more automated procedures is shifting the focus towards more advanced technological solutions [7, 8].

As literature demonstrates, the implementation of ML methods in fire risk assessment appears to be rather extensive. Such methods are largely depended on the quality and quantity of available data [10], especially in research fields such as the wildfire risk,

which is driven by interactive factors that can affect the outcomes produced. Many recent studies have advocated the efficiency of ML methods by highlighting their geospatial predictive capacity; whereas other studies undergo comparative analyses among such methods [12]. ML methods are also used for elaborating on various research dimensions of wildfires, such as fire susceptibility, risk and behaviour prediction [10]. For instance, classification by means of ML methods is used in many fire risk studies in order to identify or distinguish fine prone from non-fire prone areas, due to the accuracy of results produced by them [11].

That said, this research work attempts to compare a range of ML methods as to their robustness and comparative advantages in assessing wildfire risk in WUIs. In doing so, results produced by six literature-driven *Individual Learning Methods* [Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Logistic Regression (LR), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP)] and one *Integrated Ensemble Method* (Voting Ensemble) are compared, using a set of evaluation metrics. In addition, GIS technologies are an integral part of the spatial analysis and mapping. Implementation of ML methods uses as case study area a representative Mediterranean urbanized landscape, namely the Attica region, Greece.

The structure of the paper has as follows: Sect. 2 elaborates on the methodological framework; features aspects of the study area and establishes the dataset feeding the ML methods; and presents the theoretical ground of these methods and the metrics used for assessing their performance. Section 3 presents results obtained from the implementation of the ML methods and their performance, based on relevant evaluation metrics. Finally, in Sect. 4, critical discussion on the outcomes obtained is provided, commenting on the performance and predictive results of each single ML method in both the study area in general and the WUIs in particular.

2 Materials, Methods and Data

2.1 The Methodological Approach

The stages of the methodological approach have as follows (Fig. 1):

- *Stage I* aims at: gaining insight into the attributes of the study area; overviewing wildfire incidents and burned areas in the time span of 2000–2024; exploring literature as to the WUIs' fire-related factors; and establishing the dataset of variables relevant to the study area for implementation in the fire risk assessment methods.
- *Stage II* elaborates on: data acquisition, description and processing; short theoretical discussion on Machine Learning Methods; metrics used in the evaluation process; and selection of ML methods to be tested in Stage III.
- *Stage III* involves: overview of the dataset; implementation of the ML methods in the study area; mapping/comparison of fire risk estimations produced by the ML methods in the WUIs inherent in the study area; evaluation of results based on metrics' assessment; and validation of results through their comparison with burned areas due to wildfire incidents, taking place in the time span 2000–2024.

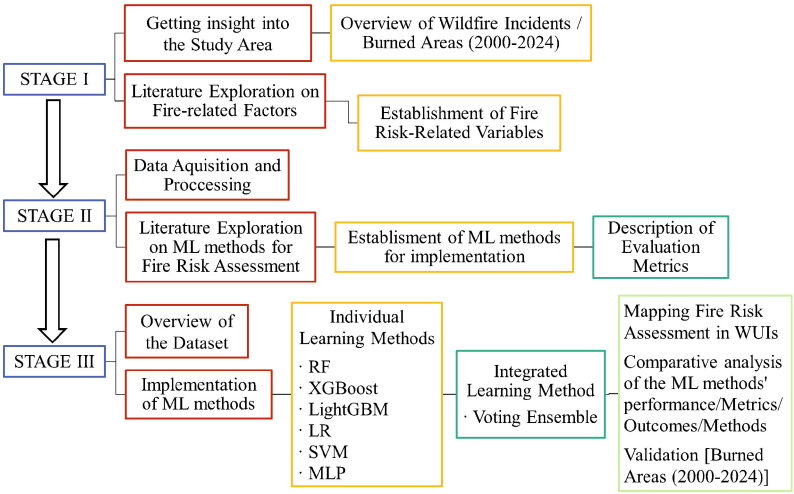


Fig. 1. Stages of the methodological approach

2.2 Delving into the Study Area for Dataset Establishment

Attica Region (Fig. 2) is selected as the study area of this work due to the recurring fire incidents, occurring in this area in the recent past. The region consists of eight Regional Units and sixty-six Municipalities. It covers an area of 3808 km^2 and, according to the 2021 Census of the Hellenic Statistical Authority (ELSTAT), gathers a population of approximately 3.81 million inhabitants [13]. The mountainous complexes prevail in the region's landscape, with the mountains of Geraneia, Pateras and Kithairon defining the western border of the region; and the mountains of Poikilo, Aigaleo, Parnitha, Penteli and Hymettus encircling the Attica basin from west to east. Furthermore, the south/south-western seaside area is washed by the Saronic Gulf, whereas the eastern/north-eastern seaside area from the South Euboean Gulf and the Gulf of Petalioi, both favouring mass tourist development of related coastal areas. The climate of the region is characterized as maritime, with dry, very warm summers [14].

A critical attribute of Attica region is the extensive urban sprawl towards the mountainous parts. This is due to the population growth in peri-urban areas, coupled with the seasonal population increase in the coastal part during summertime, i.e. from May to October, a period that coincides with the fire season. Due to the region's topographical diversity and the noticeable urban sprawl, extended WUIs are inherent in this area, i.e. urban and peri-urban areas highly exposed to wildfire risk due to their proximity to flammable vegetation [2]. The continuously expanding urban sprawl, the shift towards touristic activities in the extended coastal part and the abandonment of traditional agriculture leave the region vulnerable to wildfires and highlight the need for fire risk assessment that can properly feed proactive fire-preventing strategies [7].

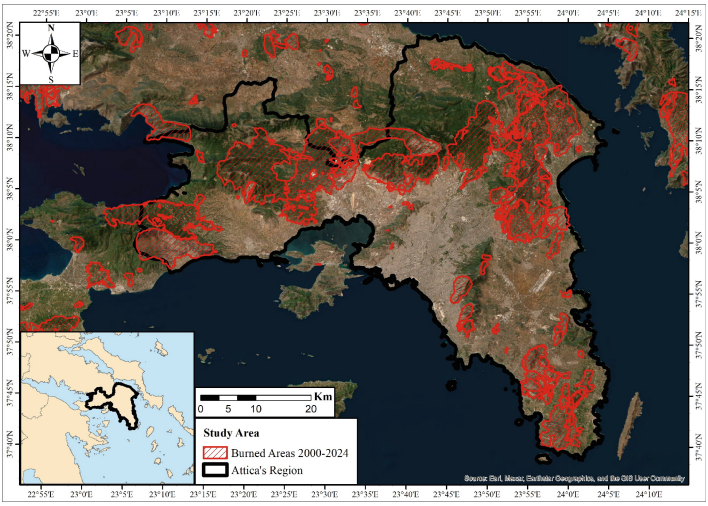


Fig. 2. Location of the Attica Region – Burned Areas in the time span 2000–2024. Source: Own Elaboration based on data from [13] and [16]

Indeed, data from the Hellenic Fire Service indicate a high frequency of fire incidents since 2000, remarkably impacting both the natural – extensive burned areas across the region, leading to the degradation of local ecosystems– and the urban environment and infrastructures, affecting thus quality of life of local population and leading to severe human losses [7]. In Fig. 3 the number of incidents occurred in this region in the time span 2000–2024 is demonstrated, coupled with the surface of burned areas [15]. Worth noticing in Fig. 3 is that although the number of incidents shows a decreasing pattern, the burned areas display an opposite one. This unveils the severity of recent wildfires, leading to extensive landscape degradation. Despite the intensive and ongoing efforts for early response and the recurrent reinforcement of fire fighting forces, the intensity of these phenomena and the size of the affected areas witness the imperative need for integrating disaster management practises in proactive strategies in order for effectively coping with today's urgencies [7].

2.3 Fire Risk Assessment Factors and Variables

The selection of fire risk related factors, as literature unveils, is a crucial task directly linked to each single study area and its particularities. Thus, choice of the most critical factors in each single case study, coupled with the successful collection of relevant data forms the ground for grasping interrelationships among these factors. This, in turn, provides important information regarding fire occurrence [8] and directly contributes to accurate and successful outcomes [17].

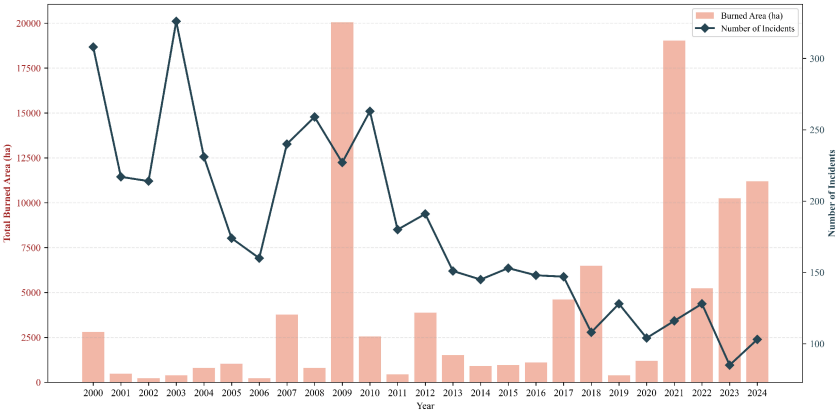


Fig. 3. Dual-axis temporal analysis of the frequency (N) of wildfire incidents and the cumulative Burned Areas (ha) in Attica Region during the Fire Season (May – October) for the time span of 2000–2024 – Data Source: Hellenic Fire Service [15]

Aksoy et al. [11] have studied a significant number of fire risk related studies and provided an extensive variables' list. Out of this list, scientists are able to select the most suitable ones, based on the specific attributes of the area at hand [7, 17]. These variables are of use regardless of the methodological approach, followed by the various studies; and are roughly categorized under a range of factors, namely Topography, Climate, Environment and Urban Space, used also in the context of this research.

Topography is closely associated with variables that refer to area's topographical and morphological attributes. It includes variables such as 'Elevation', 'Slope' and 'Aspect', which affect the behavior and progress of a forest fire, while play a decisive role in terms of accessibility to fire events and response rate of the fire fighting forces [7]. Specifically, 'Elevation' correlates with the climatic conditions of the area and the fire behaviour, whereas 'Slope' affects the speed and spread of forest fire, with steep slopes displaying a greater impact to fire spreading [18]. As for 'Aspect', the percentage of sunlight that an area receives increases the propensity of fire occurrences, making the south-southwest aspects of the study area more susceptible [19].

Climate consists of variables, such as 'Temperature', 'Wind Speed', 'Relative Humidity' and 'Precipitation', which relate to fire weather conditions [3]. Climatic conditions mostly affect the state of the environment and vegetation, both playing a decisive role in fire ignition and progress. High 'Temperatures' as well as extreme heat waves during summer time and fire season affect the vegetation's moisture, leading to higher flammability [19]; whereas low rates of 'Precipitation' affect soil moisture and area's 'Relative Humidity' [19]. The combination of decreased 'Precipitation' with high 'Temperatures' affects area's vulnerability by increasing fuel load availability and favouring fire ignition conditions [20]. Furthermore, 'Wind Speed' contributes to forest fire spreading by enhancing the fire speed and hindering fire suppression [7, 18].

Environment includes variables associated with the type, density and health of vegetation, all of decisive role in its general flammability. Specifically, 'Dominant Leaf Type' [DLT], 'Tree Cover Density' [TCD] and 'Normalized Difference Vegetation Index'

[NDVI] are the most commonly used variables in fire risk studies as they impact fire ignition and progress, i.e., significant factors to a forest fire event. 'Dominant Leaf Type' is a crucial variable since, according to various studies, coniferous trees due to their leaf structure and pine needles are significantly more prone to fire ignition in comparison to broad-leaved trees [18]. Respectively, 'Tree Cover Density' determines availability of the fuel load [18]. Thus denser vegetation in an area leads to rising fire risk [7, 11]. In recent studies, NDVI has been identified as one of the most influential fire risk variables, as vegetation's health is associated with fuel availability and vegetation's fire susceptibility [20]. NDVI provides information about vegetation's health condition. Higher values show healthy vegetation, whereas lower ones correspond to dry vegetation with higher fire risk [7, 19].

Urban Space is the factor that correlates with the urban environment and neighbouring population; and highlights the intricate involvement of the human factor in fire occurrences. As literature demonstrates, human negligence is one of the main reasons of forest fire ignition, with the number of incidents indicating a rising trend [4]. Some of the most widely used variables in forest fire research endeavours are 'Distance from Road Network', 'Distance from Rivers', 'Distance from Settlements', 'Population Distribution' and 'Land Use'. Distance variables are closely related to fire ignition as proximity of human activities to fire prone vegetation correlates with increased susceptibility [18]. In particular, proximity of WUIs' to forested areas constitutes the main reason for their high risk of ignition, being further amplified by dense population and poorly planned urban sprawl [7, 20].

Taking into consideration the particularities of the study area and the WUIs, *14 independent variables* are selected for establishing the dataset to be used in fire risk assessment (Table 1). These are: Elevation, Slope, Aspect, TCD, NDVI, Temperature (maximum), Wind Speed, Precipitation, Humidity, Distance from Road Network, Distance from Settlements, Population Distribution, Distance from Rivers and Land Use, and represent all four aforementioned factors.

2.4 Machine Learning Methods - Theoretical Background and Usability

Within the scope of this study, six *Individual Learning Methods* [five Machine Learning – Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Logistic Regression (LR) and Support Vector Machine (SVM) – and one Deep Learning (Multi-Layer Perceptron (MLP)] and one *Integrated Ensemble Method* [Voting Ensemble (Voting Classifier)] have been used to assess fire risk in the study area; and are shortly presented in the following paragraphs.

Table 1. Data and Variables used as input in the selected ML methods

Raw Data	Factor	Variables	Source	Time Span
Digital Elevation Model (DEM)	Topography	Elevation (m) Slope (%) Aspect (°)	Copernicus	-
Tree Cover Density (TCD)	Vegetation	TCD (%)	Copernicus	2023
Normalized Difference Vegetation Index [NDVI]		NDVI	Copernicus	2024
Temperature Wind Speed Precipitation Relative Humidity	Climate	Temperature (° C) Wind Speed (m/s) Precipitation (mm/y) Humidity (%)	E-OBS, Copernicus	2000-2024 Fire Season (May – Oct)
Population	Urban Space	Population Distribution (Nr)	ELSTAT	2021
Settlements		Distance from Settlements (m)	ELSTAT	2021
CLCplus Backbone		Land Use	Copernicus	2023
Road Network		Distance from Road Network (m)	Geofabrik [21]	2026
Hydrographic Network		Distance from Rivers (m)	Geofabrik	2026
Burned Areas	-	Fire/No Fire (Dependent Variable)	EFFIS, Copernicus	2000-2024

Random Forest (RF), developed by Breiman and Cutler [22], is a widely used ML algorithm. RF can handle both classification and regression problems by means of either categorical or continuous variables. It builds multiple decision trees at the same time, utilizing a "parallel ensembling" approach. By use of the Bootstrap Aggregation (Bagging) technique, every decision tree uses a different subset of the data's sample for training, while results occur from majority voting (for classification tasks) or averaging (for regression analysis). This significantly reduces variance, contributes to the avoidance of overfitting and enhances the prediction accuracy. Unlike simple decision trees, RF's robustness is grounded in the number of decision trees that are used to provide stable and efficient outcomes [10, 23].

eXtreme Gradient Boosting (XGBoost) is a powerful ML model, introduced by Chen and Guestrin [24], and is recognized for its high performance and versatility [25]. It is applicable for conducting both classification and regression tasks. Its high predictability stems from the gradient boosting framework, which iteratively combines predictions from multiple weak learners (Decision Trees). It constructs new Decision Trees sequentially, and provides more advanced ones by accounting for and minimizing the residuals of the preceding ones, thus targeting the most optimal outcomes [11, 23, 25]. Furthermore, its robustness and computational speed, as well as its regularization capability and proficiency in handling large scale and complex datasets strengthens its appropriateness to various research applications [25].

Light Gradient Boosting Machine (LightGBM) is an advanced framework, developed by Microsoft in 2017 [25, 26]. A key feature of the method is the leaf-wise tree

growth strategy, in contrast to the level-wise tree development, used by other tree-based models. LightGBM also utilizes a gradient-based one-side sampling (GOSS) technique, focusing on data instances with larger gradients (higher error), thereby enhancing models' efficiency, accuracy and training speed. Moreover, due to its speed and robustness of execution, it can handle large datasets and is applicable to both classification and regression research problems [25].

Logistic Regression (LR) is a probabilistic statistical ML model, primarily used for classification tasks. LR employs a logistic function to map the dependent variable to a probability range between 0 and 1 and is more effective when the data can be linearly separable. However, the model is susceptible to overfitting with complex data. The assumption of a linearity between input and output data can be a LR's limitation to model complex, non-linear relations between data [23].

Support Vector Machine (SVM) is a supervised ML model, applicable to classification, regression and other research applications. As a classifier, its primary objective is to define the optimal hyperplane (decision boundary) within high dimensional data, aiming to achieve the greatest distance (margin) between the hyperplane and the nearest training data point of each class, known as support vectors. In doing so, it leads to the best class separation and the lowest generalization error. In contrast to Deep Learning models that require a large amount of training samples, SVMs are more efficient and robust with limited number of training samples [10, 23].

Multi-Layer Perceptron (MLP) is a foundational Artificial Neural Network (ANN) model and a key component of Deep Learning (DL). For DL models, performance is data-driven, as both the efficiency and predictive accuracy rises with larger datasets [23]. The basic unit of a MLP are the neurons (nodes), with their architecture consisting of input, hidden and output layers. Using a feed-forward approach, every neuron of a layer connects with the subsequent one with an adjustable weight. The model reaches to the final weighted sum output layer through activation functions in order to identify complex, non-linear patterns. A significant difference with other models is that MLP uses the "Backpropagation" technique, which provides the ability to compare the predicted values to the actual ones and iteratively adjust the weights to optimize predictive accuracy [10, 23].

Finally, *Voting Ensemble (Voting Classifier)* is a ML method that aggregates the individual base learners' predictions in order to determine a prediction for the dependent variable. There are two voting strategies, namely the Hard Voting, with its prediction occurring from the majority voting of the class labels; and the Soft Voting, with its result occurring from the average of the predicted probabilities of all models. The combination of the outputs of multiple models aims at reducing variance error of individual ones and enhance predictive accuracy and robustness [27].

With regard to the evaluation of performance of the aforementioned models – accuracy and reliability – a variety of metrics can be employed. As literature demonstrates, the most commonly ones are: 'Accuracy', presenting the model's overall predictive capability and being defined as the ratio of correctly predicted instances to the total number of instances (Eq. 1) [25]; 'Precision', reflecting the model's reliability by calculating the ratio of the true positive predictions to the total number of positive predictions (Eq. 2) [25]; 'Recall', referring to the model's ability to correctly identify and predict actual

positive instances (Eq. 3) [25]; 'F1 Score', representing a balanced metric of model's performance and being the harmonic mean of Precision and Recall, ranging from 0 to 1 (Eq. 4) [27]; 'MCC' (Matthews Correlation Coefficient), being a robust statistical measure that is used to evaluate the model's binary classification quality and is particularly effective for overcoming potential issues of imbalanced datasets with values ranging from −1 to +1 (Eq. 5) [28].

The Eqs. 1 to 5 are given below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

where TP: True Positive Predictions, TN: True Negative Predictions, FP: False Positive Predictions and FN: False Negative Predictions, according to the confusion matrix categories.

Additionally, the AUC – ROC (Area Under the Receiver Operating Characteristic curve) works as a measure of the model's classification performance; the ROC curve plots the True Positive Rate [TPR (TP/(TP + FN) - correctly identified fire instances] against the False Positive Rate [FPR (FP/(FP + TN) - Non-fire instances incorrectly identified as Fire] across all possible thresholds; the AUC quantifies the overall performance within a range between 0–1 [27]; while the Spatial CV AUC (Spatial Cross-Validation AUC) assesses the model's predictive power on geographically independent data, estimating spatial generalization. This is derived as the average AUC from a k-fold cross-validation process, using spatially separated folds based on geographic groups [25, 27, 29].

3 Implementation of the Methodological Framework - Results

In this section, results obtained from the implementation of the methodological framework and the aforementioned ML methods are presented. The 14 independent variables, used as input in the fire risk assessment methods, are presented in Table 1. By use of the Burned Areas dataset for the period 2000–2024 [16], provided by the European Forest Fire Information System (EFFIS), fire occurrence points were derived and were supplemented with non-fire points (unburned areas) in order for the binary dependent variable (Fire = 1; No Fire = 0) to be established. This variable consisted of 475 points, with 222 of them being classified as 'Fire' (46.7%) and 253 as 'No Fire' (53.3%); thus, the dataset was interpreted as relatively balanced for the modelling process. All necessary

preprocessing procedures and modifications were conducted in order for the dataset to be unified and be appropriate for implementation by use of the ML methods, resulting in the final *fire risk maps*. Implementation of the selected ML methods was carried out in Python, whereas the spatial analysis and the fire risk maps were developed by use of the ArcGIS Pro 3.3.

In order for a comprehensive overview of the dataset, used in the empirical part, to be provided, descriptive statistics for all independent variables are presented in Tables 2 and 3. For the continuous variables of the dataset (Table 2), the calculated statistics are Mean value, Standard Deviation (Std), Maximum and Minimum value; while for the categorical one – i.e., Land Use – the distribution of land use classes was calculated (Table 3). As evident, the Sealed and Permanent Herbaceous categories, which correspond to urban space and permanent vegetation, are the ones that appear to have the highest frequency in the study area.

Table 2. Descriptive Statistics and Multicollinearity Assessment (VIF) for the 14 Independent Variables

Independent Variables	Count (N)	Mean	Std	Min	Max	VIF
Elevation (m)	475	239.83	223.45	0.00	1115.87	3.469
Slope (%)	475	13.39	13.91	0.00	72.94	2.018
Aspect (^{o})	475	162.89	92.33	−1.00	357.05	1.119
TCD (%)	475	9.95	19.80	0.00	87.00	2.216
NDVI	475	0.27	0.17	−0.49	0.80	1.819
Temperature (^{o} C)	475	27.80	0.78	25.84	29.80	3.150
Wind Speed (m/s)	475	3.14	0.51	2.41	4.91	4.486
Precipitation (mm/y)	475	271.15	17.42	231.44	299.78	7.820
Humidity (%)	475	57.36	1.45	53.25	59.65	5.024
Distance from Road Network (m)	475	89.14	134.32	0.00	1037.11	1.489
Distance from Settlements (m)	475	1710.97	1129.48	0.00	5393.72	1.342
Population Distribution (Nr)	475	12114.25	16723.88	148.72	64779.15	3.451
Distance from Rivers (m)	475	796.55	806.51	0.00	5803.46	1.253
Land Use	475	–	–	–	–	1.570

Table 3. Frequency Distribution and Percentage of Land Use Categories

Land Use Category	Frequency (N)	Percentage (%)
Sealed	155	32.64
Permanent Herbaceous	101	21.26
Low-growing Woody Plants	85	17.89
Woody Needle Leaved Trees	62	13.05
Non and Sparsely Vegetated	36	7.58
Woody Broadleaved Deciduous Trees	14	2.95
Woody Broadleaved Evergreen Trees	10	2.11
Periodically Herbaceous	8	1.68
Water	3	0.63
Coastal Seawater Buffer	1	0.21
Total	**475**	**100**

To ensure strong *statistical soundness* and *predictive accuracy*, Multicollinearity and Correlation Analysis was conducted. More specifically, Variance Inflation Factor (VIF) was calculated in order for multicollinearity among all independent variables to be examined and the final dataset for implementation to be determined. According to the literature, variables with VIF values greater than 10 should be eliminated due to high collinearity and redundancy [30]. As shown in Table 2, however, VIF values for all independent variables remained below the aforementioned threshold, ranging from 1.119 to 7.820. Furthermore, Spearman's rank correlation analysis was conducted to examine pairwise relationships among the continuous independent variables (Fig. 4). All variable pairs maintained correlation coefficient values within the range of 0.80 to −0.80, values which, as literature states, constitute the critical elimination thresholds. Thus the final dataset for further implementation displayed no threats or inefficiencies that could compromise produced outcomes [30, 31].

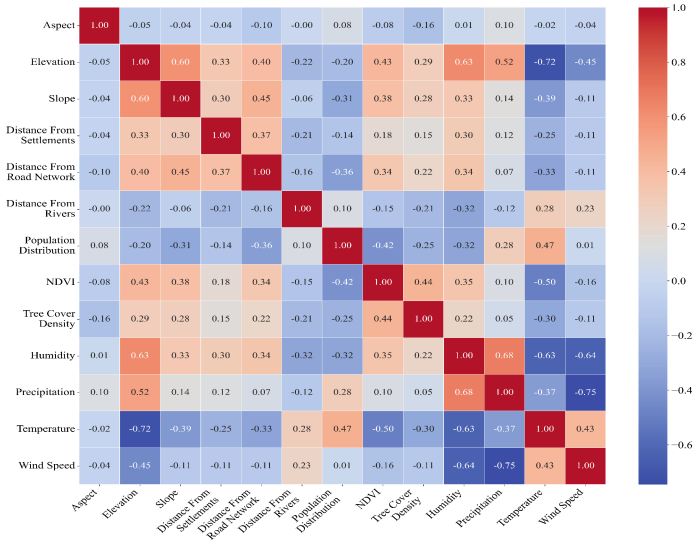


Fig. 4. Spearman Correlation Matrix for the continuous independent variables of the dataset

The aforementioned ML methods were implemented to derive the final *Forest Fire Risk Assessment maps.* The dataset was randomly partitioned into training (70%) and testing (30%) subsets. Models' training was performed using a 10-fold Spatial Cross-Validation and a GroupKFold technique in order to ensure geographical independence of the folds and avoid spatial overfitting. As for the Integrated Ensemble method, Soft Voting was implemented in this study.

Speaking of the ML methods' performance, the evaluation metrics presented in Table 4 show that all methods display high predictive accuracy and robustness. XGBoost (0.860) and Voting Ensemble (0.860) achieved the highest overall accuracy. These are followed by LightGBM (0.853), RF (0.839) and LR (0.839), while the remaining methods appear to have a slightly more conservative performance. Regarding quality achievements, XGBoost achieved the highest values in MCC (0.720) and F1 Score (0.853), indicating a robust balance between precision and recall; while MLP and SVM have the lowest values among all methods. Inspection of Test AUC values, ranging between 0.904 to 0.936, and the ROC curves (Fig. 5), verifies that all methods show strong classification performance and high capability in distinguishing between fire and non-fire instances. As for the Spatial CV AUC, Voting Ensemble (0.927), RF (0.924) and LR (0.924) exhibit the most robust performance regarding spatial generalization capabilities, slightly outperforming SVM (0.915) and XGBoost (0.913). Conversely, LightGBM (0.905) and MLP (0.906) appear to have the lowest comparative performance among all ML methods.

Table 4. Performance Evaluation Metrics for the ML Methods used in this work

Method	Accuracy	Precision	Recall	F1 Score	MCC	Test AUC	Spatial CV AUC
RF	0.839	0.833	0.821	0.827	0.677	0.930	0.924
XGBoost	0.860	0.841	0.866	0.853	0.720	0.936	0.913
LightGBM	0.853	0.811	0.896	0.851	0.710	0.926	0.905
LR	0.839	0.855	0.791	0.822	0.677	0.935	0.924
SVM	0.832	0.831	0.806	0.818	0.663	0.904	0.915
MLP	0.818	0.815	0.791	0.803	0.635	0.922	0.906
Voting	0.860	0.851	0.851	0.851	0.719	0.933	0.927

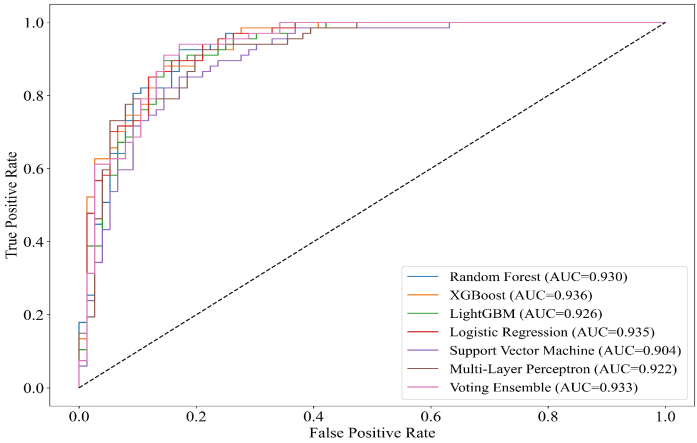


Fig. 5. ROC curve comparison and AUC values for all implemented methods

Furthermore, maps for the 14 independent variables (Fig. 6) as well as the Final Forest Fire Risk Assessment maps (Fig. 7), displaying the probability and the spatial pattern of Fire Risk in the study area, are generated. The latter are demonstrated for all methods and are coupled with Burned Areas for the time span of 2000–2024.

Worth noticing is that in the Forest Fire Risk maps, generated by the ML methods tested in this work (Fig. 7), the fire risk probability values indicate a similar range across all methods. RF (Fig. 7a) and Voting Ensemble (Fig. 7g) display a peak with a value of 0.97 and 0.98 respectively, whereas LR (Fig. 7d) reaches the maximum value of 1.0. Spatial patterns of fire risk appear to be consistent, with RF and MLP's patterns being slightly different (Fig. 7a and 7f respectively). In all ML methods, forested areas across the study area appear to have very high fire risk (brown color), while urban areas seem to have low fire risk (blue color). Areas with moderate to high fire risk (green to yellow color) appear to be at the mountain foothills, marking the transition to the urban space, i.e., Wildland Urban Interfaces (WUIs).

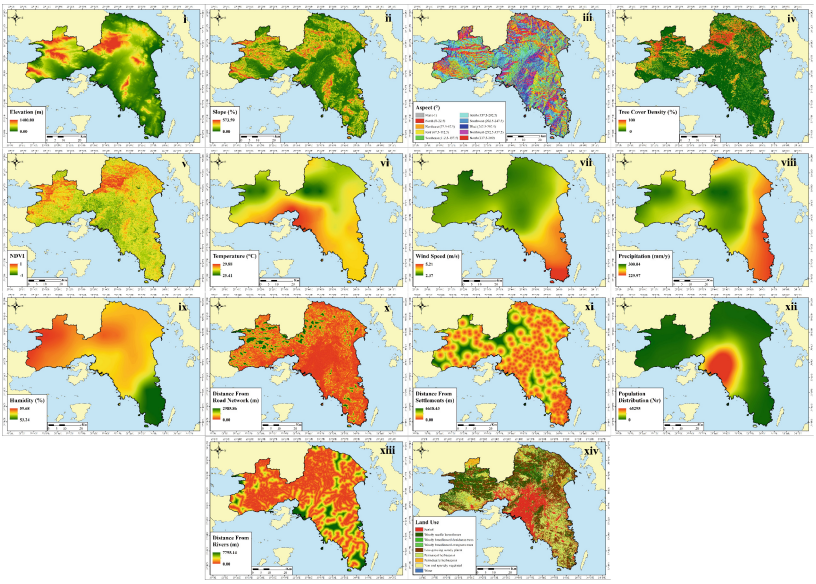


Fig. 6. Maps of the 14 independent variables: i) Elevation, ii) Slope, iii) Aspect, iv) TCD, v) NDVI, vi) Temperature, vii) Wind Speed, viii) Precipitation, ix) Humidity, x) Distance from Road Network, xi) Distance from Settlements, xii) Population Distribution, xiii) Distance from Rivers and xiv) Land Use. Source: Own elaboration.

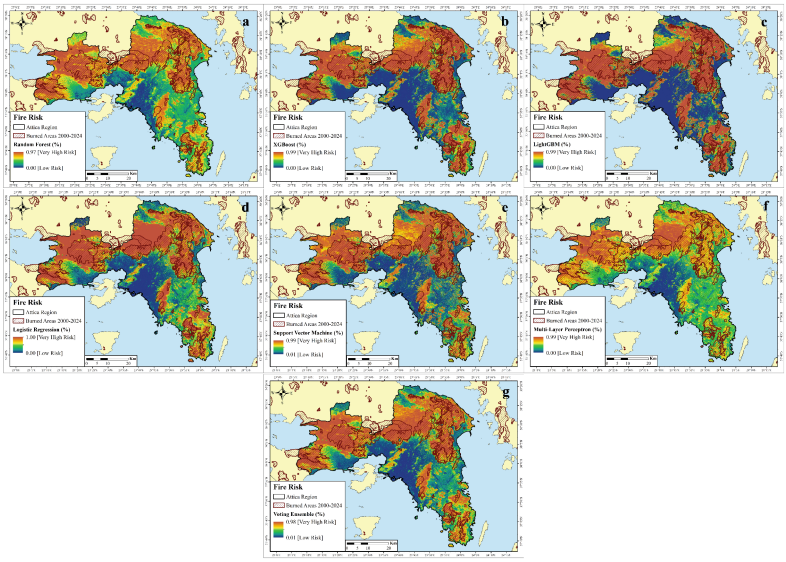


Fig. 7. Forest Fire Risk Maps produced by the selected ML methods – a) RF, b) XGBoost, c) LightGBM, d) LR, e) SVM, f) MLP and g) Voting Ensemble. Source: Own elaboration. (Color figure online)

4 Discussion and Conclusions

Vulnerability of Wildland Urban Interfaces (WUIs) and their exposure to high fire risk seems to go hand in hand with the rising number of wildfire incidents, witnessed in such areas around the globe. Mediterranean Region, as a noticeable "Hotspot" in CC terms [1], is the 'theater' of a large number of such WUIs' fire incidents, being highly noticeable due to the rising intensity, frequency and destructive outcomes in the natural and manmade environment. These highlight the need for reshaping urban planning directions in general as well as fire management strategies in particular, especially in WUIs, with a focus on a more proactive direction [4]. Of critical importance, in this respect, is to delve into efficient fire risk assessment methodologies in support of more informed policy articulation [7]. ML methods for fire risk assessment is an important improvement in fire disaster management, since state-of-the-art technologies are capable of successfully handling the complexity of this research topic [9].

That said, this study conducts a comparative analysis of fire risk assessment performance among seven ML methods, mapping respective outcomes in WUIs located in the Attica region, Greece. Results obtained as to the spatial patterns of Forest Fire Risk (Fig. 7) exhibit a high degree of consistency across all ML methods, with the highest risk values (brown color) appearing in mountainous areas, i.e. areas covered by forests and permanent vegetation, reflecting the influence of vegetation and topography in fire risk vulnerability and exposure. The lowest risk values (blue color) in all models appear to cover the urban fabric areas. The spatial patterns of XGBoost, LightGBM, LR and SVM (Fig. 7b to 7e) demonstrate more well-defined boundaries around the very high risk areas, while LR and SVM (Fig. 7d to 7e) result in minor differentiation of fire risk pattern, especially in the southern part of the study area. On the other hand, RF and MLP (Fig. 7a and 7f) demonstrate a more gradual fire risk distribution from urban to forested space, indicating the rising vulnerability of the transitional urban zones that are closely related to WUIs' distribution across the Attica region as a whole. For the study area in particular, where the WUIs consist of a large part of the region, those two methods appear to identify the area's particularities more efficiently, highlighting the important role of WUIs in the region's vulnerability and exposure, even though MLP demonstrates a slightly more conservative performance. Voting Ensemble represents an aggregate pattern, smoothing the variances of all methods. An additional strength of the high predictive performance of the ML methods is the identification of vegetated low hills within the Attica basin as high risk areas, reflecting a high level of geographical sensitivity. Fire risk assessment of all methods appears to be in alignment with the past burned areas (time period 2000–2024) (Fig. 7), a fact that serves as a qualitative validation of the methods' capability to capture the particular environmental conditions that affect fire occurrence.

Comparative analysis by means of the evaluation metrics shows that gradient boosting methods (XGBoost, LightGBM) display the highest predictive efficacy in classification tasks, with XGBoost achieving a superior level in overall predictive accuracy (0.860). Although RF performed slightly more conservatively than the gradient boosting methods, it demonstrated strong predictive capability, establishing the tree-based methods as the most efficient ones in capturing the complex, non-linear patterns of fire occurrence. In contrast, SVM and MLP demonstrate lower classification performance

(0.832 and 0.818 respectively). An important finding is the minimal divergence between Test AUC and Spatial CV AUC (ranging from −0.011 to 0.023). This indicates high spatial stability, generalization capability and robustness against spatial autocorrelation. Specifically, Voting Ensemble (0.927), RF (0.924) and LR (0.924) show the most significant performance, a fact that suggests that they are less prone to spatial overfitting and more capable of identifying the drivers that stay consistent across different geographic parts of the region. In contrast, the lower stability of LightGBM (0.905) and MLP (0.906) suggests less spatial flexibility. The Voting Ensemble represents a balanced approach as it mitigates individual variances and weaknesses, without compromising the overall predictive performance and geographical reliability. Overall, all methods demonstrate superior performance, stressing the suitability of tree-based methods (XGBoost, LightGBM and RF) as frontrunners for fire risk assessment and fire-related research.

The efficacy of ML frameworks for informing proactive fire-related policy decisions is tightly interwoven with the quality and quantity of training data that affects reliability of outputs. Limited datasets often influence the methods' ability for generalization, leading to potential overfitting and weak statistical stability. On the contrary, ML methods require larger volumes and high quality of data in order to identify complex patterns, making the selection and pre-processing of the dataset an important determinant of their performance. Moreover, the integration of time variant data (e.g. meteorological data) alongside with static variables can provide more up-to-date fire risk outcomes and enhance preparedness.

In concluding, the aforementioned analysis leads to robust results as to the fire risk assessment in the specific study area, witnessed by the rather converging outcomes of the tested methods. Such an inference forms a solid ground for planning and policy reaction, targeting vulnerability decrease of the high risk parts embedded in the study area. In addition, results demonstrate that such parts seem to be strongly related to the impact of poorly-planned urban sprawl in flammable peri-urban zones, clearly demonstrating the vulnerability of the WUIs and the necessity for dedicated proactive urban planning strategies, especially in those parts. Indicative policy interventions may range from the establishment of low-flammability green zones and firebreaks to land use changes/constraints, risk zoning, awareness raising of local population, preparation of evacuation plans, etc., all contributing to the reduction of fire risk and fuel continuity and assisting firefighting intervention, In addition, fire risk assessment maps could provide good insight for effectively allocating firefighting resources that may enhance response rate of fire forces in cases of fire emergency.

Acknowledgments. This research was conducted as part of the Doctoral Dissertation of Ms. Angeliki Papazoglou and was funded by the Special Account for Research Funding (E.L.K.E) of the National Technical University of Athens (NTUA).

References

1. IPCC, 2023: Climate Change 2023 - Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change

[Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland, Intergovernmental Panel on Climate Change (IPCC) (2023). https://doi.org/10.59327/IPCC/AR6-9789291691647. Accessed 12 Feb 2026
2. Ganteaume, A., Barbero, R., Jappiot, M., Maillé, E.: Understanding future changes to fires in southern Europe and their impacts on the Wildland-Urban Interface. J. Saf. Sci. Resil. **2**(1), 20–29 (2021). https://doi.org/10.1016/j.jnlssr.2021.01.001
3. Papavasileiou, G., Giannaros, T.M.: Synoptic-scale drivers of fire weather in Greece. Sci. Total Environ. **925**, 171715 (2024). https://doi.org/10.1016/j.scitotenv.2024.171715
4. Kelley, D.I., et al.: State of Wildfires 2024–2025. Earth Syst. Sci. Data **17**(10), 5377–5488 (2025). https://doi.org/10.5194/essd-17-5377-2025
5. Gemitzi, A., Koutsias, N.: A Google Earth Engine code to estimate properties of vegetation phenology in fire affected areas – a case study in North Evia wildfire event on August 2021. Remote Sens. Appl. Soc. Environ. **26**, 100720 (2022). https://doi.org/10.1016/j.rsase.2022.100720
6. Dosiou, A., et al.: Employing Copernicus Land Service and Sentinel-2 satellite mission data to assess the spatial dynamics and distribution of the extreme forest fires of 2023 in Greece. Fire **7**(1), (2024). https://doi.org/10.3390/fire7010020
7. Papazoglou, A., Lagarias, A., Stratigea, A.: Methodologies for Wildland-Urban Interface (WUI) fire risk assessment – Case study North-Eastern Attica Region. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2025 Workshops. ICCSA 2025. LNCS, vol. 15889, pp. 88–105. Springer, Cham (2026). https://doi.org/10.1007/978-3-031-97603-2_6
8. Zhang, G., Wang, M., Liu, K.: Deep neural networks for global wildfire susceptibility modelling. Ecol. Ind. **127**, 107735 (2021). https://doi.org/10.1016/j.ecolind.2021.107735
9. Sevinç, V.: Mapping the forest fire risk zones using artificial intelligence with risk factors data. Environ. Sci. Pollut. Res. **30**, 4721–4732 (2023). https://doi.org/10.1007/s11356-022-22515-w
10. Jain, P., Coogan, S.C.P., Subramanian, S.G., Crowley, M., Taylor, S., Flannigan, M.D.: A review of machine learning applications in wildfire science and management. Environ. Rev. **28**(4), 478–505 (2020). https://doi.org/10.1139/er-2020-0019
11. Aksoy, E., Kocer, A., Yilmaz, İ., Akçal, A.N., Akpinar, K.: Assessing fire risk in Wildland–Urban Interface regions using a machine learning method and GIS data: the example of Istanbul's European side. Fire **6**(10) (2023). https://doi.org/10.3390/fire6100408
12. Wei, G., Han, G.S., Lang, X.: Fire risk assessment using machine learning techniques: a case study of Jinan City, China. Sci. Rep. **16**, 6410 (2026). https://doi.org/10.1038/s41598-026-37074-0
13. ELSTAT: Hellenic Statistical Authority. https://www.statistics.gr/en/home/. Accessed 28 Feb 2026
14. Krina, A., Koutsias, N., Pleniou, M., Xystrakis, F.: Climatic classification of Greece: update–future estimation–relation with forest vegetation. In: Proceedings of the 18th Panhellenic Forestry Conference & International Workshop, Edessa, Greece, pp. 1088–1095 (2017). (in Greek)
15. Fire Department. Hellenic Fire Service. https://www.fireservice.gr/el. Accessed 28 Feb 2026
16. Copernicus. Copernicus | Europe's eyes on Earth. https://www.copernicus.eu/en. Accessed 04 Feb 2026
17. Borisova, B., et al.: Wildfire risk assessment and mapping – an approach for Natura 2000 forest sites. Trees Forests People **16**, 100532 (2024). https://doi.org/10.1016/j.tfp.2024.100532
18. Sivrikaya, F., Küçük, Ö.: Modeling forest fire risk based on GIS-based analytical hierarchy process and statistical analysis in Mediterranean region. Eco. Inform. **68**, 101537 (2022). https://doi.org/10.1016/j.ecoinf.2021.101537

19. Jiang, W., et al.: Wildfire risk assessment using deep learning in Guangdong Province, China. Int. J. Appl. Earth Obs. Geoinf. **128**, 103750 (2024). https://doi.org/10.1016/j.jag.2024.103750
20. Parvar, Z., Saeidi, S., Mirkarimi, S.: Integrating meteorological and geospatial data for forest fire risk assessment. J. Environ. Manage. **358**, 120925 (2024). https://doi.org/10.1016/j.jenvman.2024.120925
21. Geofabrik: OSM - OpenStreetMap Data | Geofabrik Downloads. https://download.geofabrik.de/europe/greece.html. Accessed 28 Feb 2026
22. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
23. Sarker, I.H.: Machine Learning: algorithms, real-world applications and research directions. SN Comput. Sci. **2**(3), 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
24. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 785–794. Association for Computing Machinery, New York (2016). https://doi.org/10.1145/2939672.2939785
25. Tonbul, H.: Integrating ensemble machine learning and explainable AI for enhanced forest fire susceptibility analysis and risk assessment in Türkiye's Mediterranean region. Earth Sci. Inf. **17**, 5709–5731 (2024). https://doi.org/10.1007/s12145-024-01480-7
26. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 3149–3157. Curran Associates Inc., Red Hook (2017). https://dl.acm.org/doi/10.5555/3294996.3295074
27. Huang, S., Ji, J., Wang, Y., Li, W., Zheng, Y.: Development and validation of a soft voting-based model for urban fire risk prediction. Int. J. Disaster Risk Reduct. **101**, 104224 (2024). https://doi.org/10.1016/j.ijdrr.2023.104224
28. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. **21**, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7
29. Kumar, C., Walton, G., Santi, P., Luza, C.: Random Cross-Validation produces biased assessment of Machine Learning performance in regional landslide susceptibility prediction. Remote Sens. **17**(2), 213 (2025). https://doi.org/10.3390/rs17020213
30. Nur, A.S., Kim, Y.J., Lee, J.H., Lee, C.-W.: Spatial prediction of wildfire susceptibility using hybrid Machine Learning models based on Support Vector Regression in Sydney, Australia. Remote Sens. **15**(3), 760. https://doi.org/10.3390/rs15030760
31. Diaz-Vallejo, M., et al.: Analyses of the variable selection using correlation methods: an approach to the importance of statistical inferences in the modelling process. Ecol. Model. **498**, 110893 (2024). https://doi.org/10.1016/j.ecolmodel.2024.110893

SmartParcel System: A Geospatial Platform for Enhancing Cross-Fertilisation Between Spatial Planning and Territorial Cohesion Policies in Serbia

Ljiljana Živković(✉)

Republic Geodetic Authority, Belgrade, Serbia
liliana.zivkovic@gmail.com

Abstract. Building on the experience of Cohesion Policy (CP) implementation in the 2021–2027 programming period, the European Commission has proposed adjustments for 2027–2034 aimed at enhancing its flexibility and strategic alignment. Recent policy discussions emphasise the need to strengthen cross-fertilisation between CP and national spatial planning systems, particularly at the strategic (national) level, where integration capacities proved to be the highest. At the same time, the ongoing digitalisation of spatial planning practices across Europe offers new opportunities to improve policies coordination, data integration and evidence-based decision-making.

In this context, Serbia, as an EU candidate country, is advancing reform towards a more transparent, data-driven and participatory spatial planning system. As part of these efforts, the Republic Geodetic Authority, in partnership with Lantmäteriet, developed the SmartParcel system within the SIDA-supported project "ePlan4eSpace – SmartSDI". This geospatial platform integrates multiple planning-related datasets through a Serbian National Spatial Data Infrastructure (NSDI) environment, enabling improved coordination among stakeholders and supporting the alignment and integration of spatial planning with sectoral policies and future CP instruments.

The paper presents the SmartParcel system as a digital enabler of cross-fertilisation and discusses its potential to support a data-driven, place-based approach to cohesive territorial development in Serbia.

Keywords: spatial planning · territorial cohesion · geospatial platform · cross-fertilisation · SmartParcel system

1 Introduction

Responding to the progress achieved to date, as well as the identified strengths and weaknesses in the implementation of Cohesion Policy (CP) during the 2021–2027 programming period, and considering that development activities aimed at territorial cohesion are still ongoing across all Member States and the European Union (EU) as a whole, the European Commission has recently proposed several adjustments to this key "leave no one behind" development instrument for the 2027–2034 period.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 641–658, 2027.
https://doi.org/10.1007/978-3-032-30521-3_40

Among the proposed changes is a refinement of the *shared management model* between the EU and Member States, which is expected to become more flexible, results-oriented, and focused on regional partnership [4]. In addition to improvements within CP itself, a recent ESPON Policy Brief highlights the need to strengthen national sectoral policies and their territorial instruments, particularly by enhancing their cross-fertilisation with spatial planning policies, while respecting the "do no harm to cohesion" principle [11]. Compared to sectoral policies, spatial planning systems in Member States have already established well-defined legal and institutional frameworks that support such cross-fertilisation. These frameworks have also proven to be effective in facilitating stronger integration of CP instruments into the spatial planning policies, including their incorporation into national development budgets.

Furthermore, experience from previous CP implementation periods indicates that spatial planning systems have supported cross-fertilisation more effectively than other territorial cohesion policy systems, particularly at the national –strategic- level. This suggests that institutions operating at the strategic development level are generally more responsive and flexible in integrating CP instruments than local-level institutions, which tend to have more regulatory roles. In pursuit of a more integrated approach to achieving territorial cohesion across Member States, experts have therefore proposed seven steps to enhance cross-fertilisation between spatial planning and CP at the national level [7]. This also reflects an anticipated shift in the next programming period from a bottom-up and middle-up approach towards a more top-down approach in coordinating national territorial instruments.

At the same time, since 2011, spatial planning practices across Europe have increasingly focused on the digitalisation of planning data and processes, driven by the need to improve efficiency, transparency, innovation capacity and accessibility of planning information, often through the establishment of centralised plan registers and national platforms [9]. Beyond supporting *vertical* (across governance levels), *horizontal* (across sectors and stakeholders) and *territorial/jurisdiction* coordination, the digitalisation of spatial planning systems also has significant potential to enhance cross-fertilisation between CP and spatial planning at the national level [2], particularly by enabling the implementation of the identified seven improvement steps [7].

In line with the broader trend of the "Europeanisation of spatial planning" [2], driven by various territorial initiatives (such as Interreg programme, CP, EU sectoral policies, the Territorial Agenda, and ESPON) and governance processes, Serbia, as a candidate country, is undertaking reform aimed at digitalising its spatial planning system [31]. This reform seeks to address existing system weaknesses while establishing a more sustainable, transparent, and participatory planning framework capable of effectively supporting present and future EU investments. As a first step towards the so-called eSpace reform, the Republic Geodetic Authority, in partnership with Lantmäteriet, developed and launched in 2025 the SmartParcel systems within the SIDA-supported project "ePlan4eSpace – SmartSDI" [13, 14].

The SmartParcel system currently comprises three data catalogues -Catalogue of Legal Acts, Catalogue of Planning Documents, and Catalogue of Entities Accredited for Issuing Land Use Conditions- and a planned land use (PLU) database alongside a geospatial platform that enables the integrated visualisation and use of these datasets in

combination with other relevant spatial and thematic data. This data integration capability positions SmartParcel as a key geospatial platform for supporting cross-fertilisation and coordination between spatial planning, sectoral policies and planning stakeholders [15], as well as for enabling future integration with CP. Moreover, it provides the foundation for a data-driven, place-based approach to sustainable territorial development in Serbia.

The aim of this paper is to present the SmartParcel system and analyse its potential to enhance cross-fertilisation between spatial planning and other territorial cohesion policies in Serbia. Following this introduction, Sect. 2 examines the current state of CP implementation in Europe, including opportunities and mechanisms for strengthening its integration with spatial planning, as well as recent trends in the digitalisation of planning practices. Section 3 provides an overview of the current spatial planning and urban development system in Serbia. Section 4 presents the main components and functionalities of the SmartParcel platform and its role in addressing existing spatial planning system limitations. The paper concludes in Sect. 5 with key findings and directions for future development.

2 Towards Territorial Cohesion in Europe

The Cohesion Policy (CP), reformed following the Barca Report in 2009 towards a place-based approach focusing on green, digital, and social transition interventions, represents today the main EU investment strategy aimed at reducing social, economic, and territorial disparities across European regions [11]. It supports regional growth and competitiveness in coordination with national, evidence-based territorial policies, strategies, and their instruments. By mobilising EU funds through specific place-based territorial instruments (e.g. ISUD, CLLD, ITI, Interreg), CP investment decisions are directed towards achieving the five key policy objectives (PO) through projects grouped into corresponding intervention areas [3, 11].

Despite significant efforts invested in CP implementation and the delivery of targeted economic and social benefits across Member States through various territorial mechanisms and investments (Fig. 1), territorial cohesion outcomes in Europe remain uneven [7]. The literature identifies multiple reasons for these mixed results, ranging from cultural differences and variations in planning traditions, regulatory frameworks, and tools, to insufficient cooperation between national institutions responsible for CP implementation and territorial policies [6, 17]. A growing body of research emphasises that, with regard to territorial cohesion, the degree of cross-fertilisation between CP and spatial planning policies, particularly at the national level (Fig. 2), plays a decisive role in differentiating between lagging and non-lagging regions, as well as between resilient and more vulnerable territories.

Recognising the responsiveness and suitability of comprehensive national spatial planning systems and policies for enhancing cross-fertilisation with CP and other territorial instruments, a recent ESPON Policy Brief proposes seven practical steps to strengthen this interaction [7]. These recommendations are particularly relevant for countries where CP constitutes a significant share of public investment. Among them, the step "Champion (or support) joint working" is especially pertinent, as it promotes voluntary

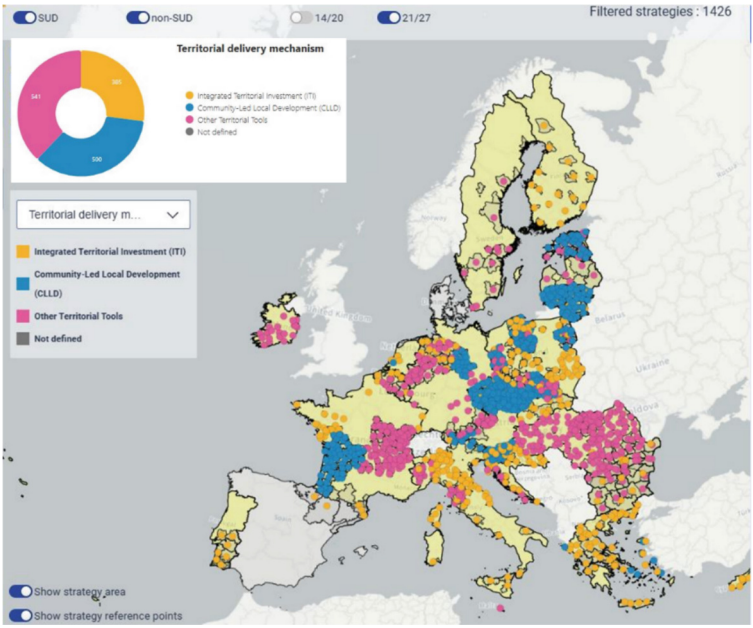


Fig. 1. EU Cohesion Policy's territorial delivery mechanisms have been applied towards the cohesive development investments in a number of strategies adopted and implemented by the Member States during 2021–2027 programming cycle [3]

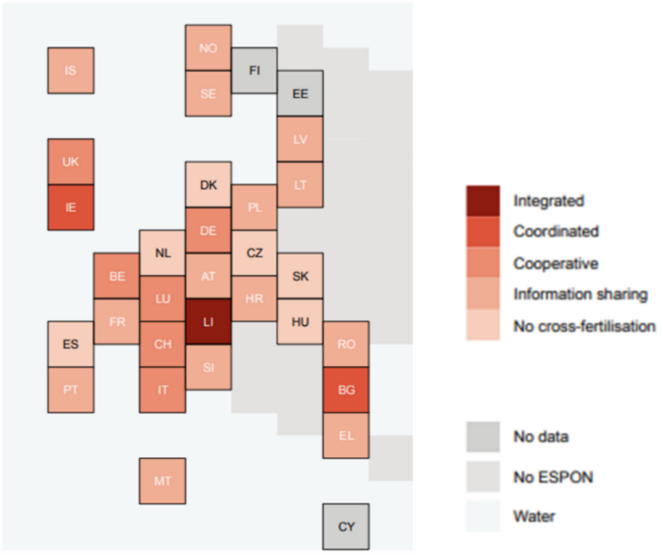


Fig. 2. Overall evaluation of cross-fertilisation between cohesion policy and spatial planning [7]

cooperation through the establishment of more formal and permanent platforms bringing together national and subnational authorities, agencies and stakeholders to guide strategic spatial planning, particularly in functional regions [1, 7].

At the same time, in pursuit of greater efficiency, innovation capacity, transparency, and integration of planning and related data, Member States have been actively advancing the digitalisation of spatial planning systems since 2011 (Fig. 3). [9] This process is grounded in common European frameworks, notably the (simplified) INSPIRE Directive and related strategies, which enable integrated access to and analysis of the spatial planning and sectoral data both within and across countries [34]. In addition to improving coordination, digitalised spatial planning systems provide a foundation for enhanced cross-fertilisation among territorial policies and instruments, including CP. Furthermore, centralised registers of planning documents and associated geospatial platforms support transparency, communication, and stakeholder participation [1, 15], while also facilitating the implementation of recommended steps for strengthening cross-fertilisation and reinforcing the conditions for future cohesive territorial development in Europe [7, 9].

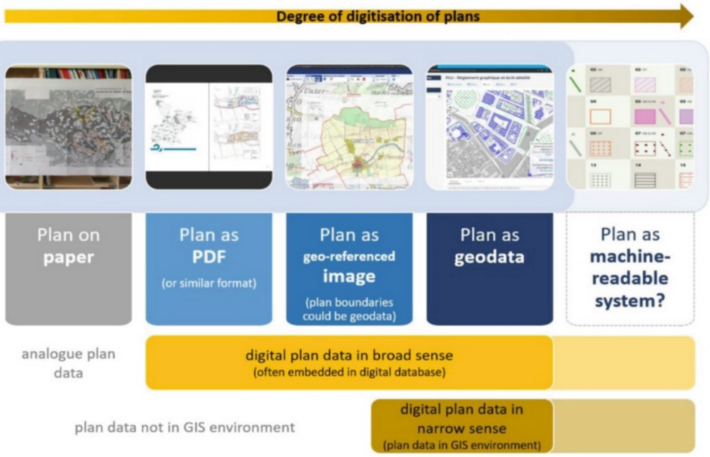


Fig. 3. Degree of digitalisation of plans in Europe: in the majority of surveyed countries plan data are in form of geodata, i.e., representation of the actual binding analogue plan [9]

3 Spatial Planning System in Serbia

Today, Serbian cities are home to approximately 60% of the country's population. However, rather than driving green, sustainable and resilient growth, they face a range of challenges, including population decline (with the exception of Novi Sad and Novi Pazar), uneven spatial development, environmental degradation and inadequate communal infrastructure and services. In addition, they are increasingly vulnerable to extreme weather events and climate change (e.g. floods, heat waves and droughts, seismic hazards, and risks associated with informally constructed buildings) [33]. At the same time, local spatial planning practices remain suboptimal, resulting in inefficient land use and limited investment capacity [8, 27–29].

In response, the Serbian Government has adopted the Sustainable Urban Development Strategy [18] and upgraded related national policies and regulations in areas such as spatial planning, public finance, service delivery and infrastructure management. These efforts have created a more favourable policy framework for environmental management and climate action, aligned with the EU Green Deal and other European and global initiatives. However, these policy objectives have not yet been sufficiently translated into concrete actions at the local level. This gap limits improvements in urban livability and performance, and constrains efforts to address inefficient land management, weak spatial planning practices and the lack of strategic, well-informed investments essential for the green transition.

To support the Government in translating national development priorities into effective local policies and investment portfolios, the World Bank, in its recent Western Balkans and Croatia Urbanization and Territorial Review, recommends the "3C" framework (Fig. 4).



Fig. 4. 3C framework for strengthening important role of cities for modern economy, sustainable development and green transition of Serbia, i.e. cohesive development [29]

Regarding local spatial planning practices, current weaknesses significantly hinder necessary reforms (e.g. achieving the "No Net Land Take by 2050" objective [5, 8, 17]). Although Serbia has a long tradition in spatial planning, its system remains complex and insufficiently coordinated [27]. It requires reform, streamlining and adaptation ("fit for purpose") to effectively respond to contemporary social, economic and environmental challenges. At the local level, the large number of mandatory and recommended strategic, sectoral and spatial planning documents places a significant burden on administrative capacities. Limited human, technical, and financial resources constrain local authorities' ability to ensure sectoral integration, cross-fertilisation, and alignment of local and national policies and investment instruments within planning documents (spatial and urban plans).

In the context of Serbia's EU accession process, these challenges also affect the integration of European CP instruments and territorial mechanisms. Capacity-building efforts are already underway in Serbia through initiatives such as the EU INTEGRA Programme [10] and the GIZ-supported COHPOL 2 project [12]. In addition, local spatial planning systems need to strengthen participatory processes and establish stronger links with strategic investment prioritisation and decision-making frameworks at both

national and EU levels, including CP, as envisaged in the draft Spatial Plan of the Republic of Serbia [19].

The World Bank also highlights the need to improve local spatial planning through enhanced data quality and the development of GIS-based systems and geospatial platforms to support inter-agency coordination (Fig. 5) [28]. Currently, geographic information systems (GIS) are used in approximately one-third of Serbian cities and municipalities, primarily for internal purposes such as plan preparation and analysis. Their use for external purposes, such as interactive data sharing with stakeholders and the public, remains limited. Although the Law on Planning and Construction [35] requires the establishment of geospatial information systems at the local level, most municipalities have not yet implemented such systems or developed centralised databases and geospatial platforms, largely due to the insufficient coordination among their sectoral departments and public enterprises.

CHALLENGE	RECOMMENDATION
PLANNING: Capacity constraints to meet urban planning obligations and to support climate and disaster-risk informed planning	• At the national level, streamline the planning system by simplifying administrative requirements for mandatory and optional planning documents, adjusting to realities on the ground (at the LSG level); • Support horizontal and vertical cooperation to ensure alignment between national and local plans (including spatial, urban development and sectoral plans); • Improve capacity of LSGs to integrate land use and sectoral plans to promote sustainable development through coordinated investments; • Develop and implement climate resilient and risk-informed urban plans (for all identified significant hazards, e.g., landslides, earth-quakes, floods for category II rivers); • Develop and maintain updated disaster risk assessments, disaster risk reduction plans and contingency plans to ensure the continuity of critical infrastructure in the event of disasters; and • Improve the quality of data, to support evidence-based planning and investment decision-making and to monitor performance indicators. This includes developing and enhancing GIS systems and developing a digital platform for inter-agency coordination to improve service planning and delivery.

Fig. 5. 5 recommendations for overcoming the current spatial planning practice challenges in Serbian cities and municipalities (i.e. local self-governments, LSGs) [28]

Establishing centralised geospatial databases and platforms would not only improve public service planning and delivery but also support cross-fertilisation and integration of policies and investment instruments across governance levels [1]. As part of ongoing efforts to strengthen investment planning capacities, Serbian cities and municipalities have identified and registered infrastructure projects under the World Bank/AFD-supported "Local Infrastructure and Institutional Development" (LIID) project, aimed at enhancing capacities for sustainable and climate-smart infrastructure development [20].

In line with broader European trends in the digitalisation of spatial planning, Serbia has also launched the so-called eSpace reform [31], which envisions the establishment of an integrated geospatial platform to support spatial planning as a transparent, participatory, and data-driven process - from plan preparation and adoption to its implementation.

This reform, conceptually introduced in 2021 and formalised through the Law on Planning and Construction in 2023, includes two main components: ePlan and eSpace [31]. The first development phase focuses on establishing a central, legally reliable register of planning documents, along with the registers of key spatial planning stakeholders and datasets related to land use and development. These registers are expected to support -in the second phase- development of a comprehensive national geospatial platform – eSpace- for place-based, data-driven decision-making and territorial policy management at both national and local levels.

4 SmartParcel System

As a first step in the reform of the spatial planning system in Serbia, and the beginning of its digitalisation in line with relevant INSPIRE thematic specifications, the recommendations of the ESPON DIGIPLAN project [9], and the eSpace reform, the Republic Geodetic Authority (RGA), in partnership with its Swedish counterpart Lantmäteriet, established the SmartParcel (or ПаметнаПарцела) system in 2025.

Being the main outcome of the SIDA-supported project "ePlan4eSpace – SmartSDI" [32], this system's concept was developed in accordance with the recommendations of the UN Framework for Effective Land Administration (FELA) [25]. The system architecture and functionalities are based on European best practices in the digitalisation of spatial planning systems and practices, while also reflecting requirements collected from Serbian cities and municipalities through an online survey conducted in late 2024 [24].

The aim of establishing the SmartParcel system is to create a geospatial platform that brings together all relevant spatial planning stakeholders in Serbia and integrates authoritative information on each cadastral parcel in a single environment. In this way, the system provides an integrated platform that supports a place-based approach to the sustainable management of land and other resources, as well as preparation and implementation of local and national plans and policies in a transparent, participatory and inclusive manner [13, 14].

The objectives of the SmartParcel system respond to the requirements of a modern spatial planning and land management system by ensuring the availability of accurate, comprehensive and reliable information for each cadastral parcel, in line with the European and international best practices and specific local needs.

In the context of rapidly evolving economic, social and environmental conditions, evidence-based decision-making supported by high-quality geospatial data has become essential. Whether for strategic projects, public or private investments, or individual decisions by landowners, users increasingly require fast and reliable access to relevant information on the current status and development potential of specific cadastral parcels and real estate assets, and this is where the SmartParcel system emerges as a enabler, providing a comprehensive, integrated and user-oriented solution.

4.1 System Concept

As already mentioned, the SmartParcel system concept is grounded in the UN FELA, which promotes the establishment of modern and efficient land administration systems

capable of supporting the implementation of the Agenda 2030 and the Sustainable Development Goals (SDGs). In this context, land administration systems are expected to integrate and provide four key types of information related to land or cadastral parcel: *value, ownership, use and development potential,* and *technical characteristics.*

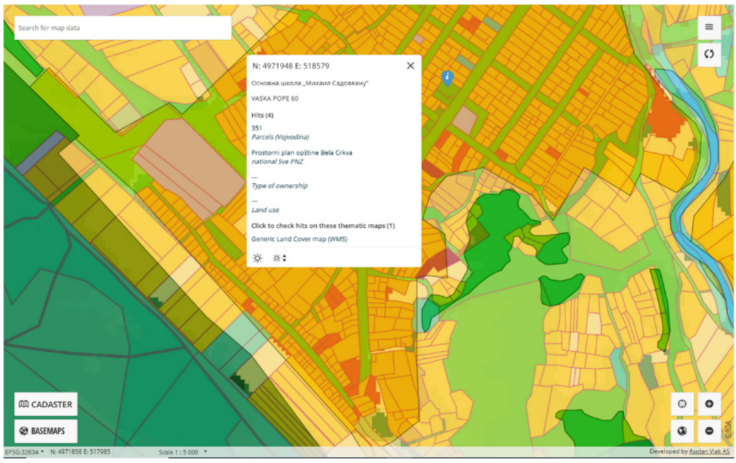


Fig. 6. SmartParcel system concept is in line with UN FELA requirements

To enhance its capabilities and evolve towards a modern land administration system that delivers authoritative public data and user-oriented services (for citizens, businesses, and public sector entities), the RGA is improving the quality and availability of existing cadastral information, such as data on *parcel value*, *ownership*, *boundaries* and *existing land use*, while also establishing the SmartParcel system to collect and provide information on *planned land use* as defined in planning documents (spatial and urban plans) (Fig. 6).

Established within the Serbian National Spatial Data Infrastructure (NSDI) [23], the SmartParcel system builds upon existing infrastructure and integrates additional RGA datasets and services to provide comprehensive information on the development potential of cadastral parcels (Fig. 7). On the production side, the system consists of a Web platform (landing page) and data collection tools supporting the three new data catalogues (No. 1) and Knowledge Base. These catalogues are populated by authorised stakeholders and made publicly accessible through NSDI distribution services, together with other relevant RGA datasets and services for sustainable land management (e.g. eCadaster, Disaster Risk Register, and other GeoSrbija registers), as well as geospatial datasets on planned LU developed during the project (No. 2).

In addition, the SmartParcel system is connected via Web services to environmental and other relevant land-use regime datasets (e.g. sectoral plans and policies data) (No. 3), which are maintained by the responsible authorities in accordance with the sector-specific regulations and the Law on NSDI.

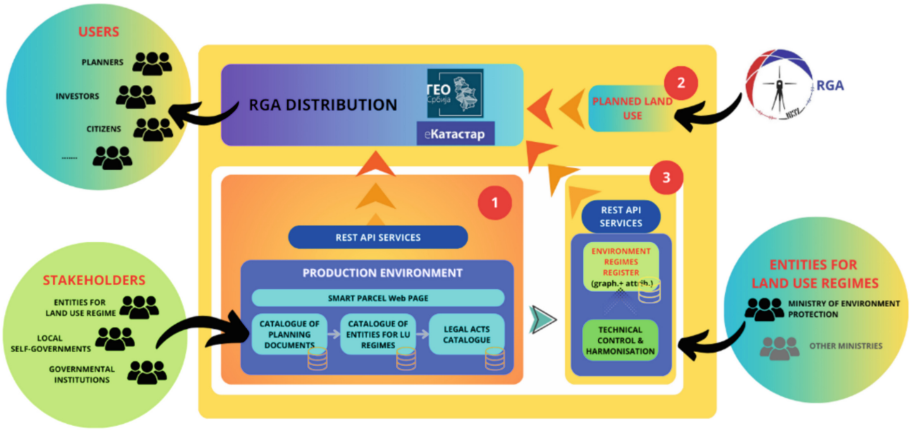


Fig. 7. SmartParcel system concept: system structure presents integration of a Web platform for data production and access; a data collection tools and a REST API services

The SmartParcel system concept reflects best practices in geospatial data management and service provision. At the same time, it fosters stakeholder ownership, which is essential for the system's long-term sustainability, scalability and further development.

4.2 System Components

The SmartParcel system consists of several key components: a Web platform (landing page), a Knowledge Base, a planned land use (PLU) database, and three data catalogues, namely the Catalogue of Planning Documents, the Catalogue of Legal Acts, and the Catalogue of Entities Accredited for Issuing Land Use Conditions (hereafter referred to as land use regimes).

While the Web platform, Knowledge Base and PLU database are publicly accessible, the access to catalogue applications is restricted to authorised stakeholders, including cities and municipalities (i.e. local self-governments), ministries and other public authorities. These users are granted controlled access to the catalogue databases within the system's production environment, where they operate under defined roles (Editor or Supervisor) to enter new data and update existing records related to planning documents, land use regimes and associated legal acts.

Catalogue of Planning Documents

The Catalogue contains key information on each spatial and urban plan (or planning document) adopted in Serbia, including their planning areas (defined as polygons representing the spatial extent covered by each plan), associated metadata, as well as their textual and graphical parts, in accordance with the Law on Planning and Construction.

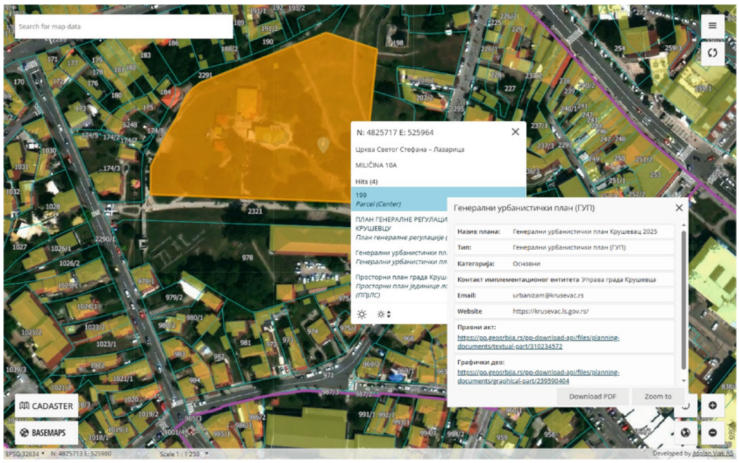


Fig. 8. Overview of data in the Catalogue of planning documents



Fig. 9. The SmartParcel system' home or landing page, and Catalogue of planning documents' editor and data collection interface (dialog boxes)

The availability of geospatial data on planning areas enables the identification of applicable plans for each cadastral parcel through simple spatial selection within the Overview map. As a result of selecting a cadastral parcel, the system generates a list of

all plans applicable to that parcel, together with access to the corresponding textual and graphical parts' documentation, which users can download and search (Fig. 8).

Access to the data collection tool of this and other Catalogues is restricted to authorised users, who can create new records or edit and update existing entries within the Catalogue database (Fig. 9).

Catalogue of Entities Accredited for Land Use Conditions Issuing

The Catalogue contains essential information on entities responsible for issuing conditions related to land use and development (e.g. nature protection, water management, cultural heritage protection and access to infrastructure systems). Similar to planning documents, the jurisdictions of these entities are spatially defined in the Catalogue through geospatial data (polygons).

This enables users of the SmartParcel system to identify, through simple spatial selection of a cadastral parcel, which public authorities—at the local and/or national level—have jurisdiction over its use, as well as the corresponding legal basis for such jurisdiction (Fig. 10).

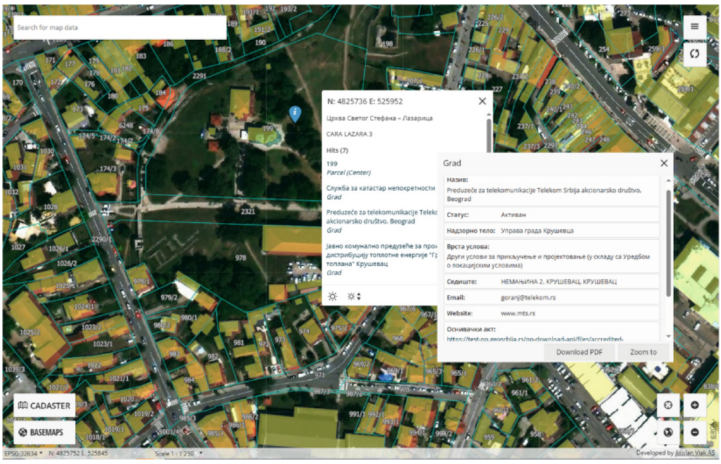


Fig. 10. Overview of data in the Catalogue of entities accredited for land use conditions issuing

Catalogue of Legal Acts

This Catalogue complements the previous two and contains all relevant laws, by-laws, decisions, and regulations that constitute the legal basis for planning documents and land use regimes applicable to a selected cadastral parcel. Each legal act is linked to the institution that enacted it, as well as to the corresponding planning document, and includes information on the official gazette in which it is published.

In this way, the system enables users to access both geospatial and legal information related to the development potential of each cadastral parcel within a single platform.

Knowledge Base

As an integral component of the SmartParcel system, the Knowledge Base provides

experts, institutional staff, investors and the general public with access to information on system functionalities, as well as professional and educational content. It also facilitates the exchange and advancement of knowledge and professional practice in the fields of spatial planning, land management and sustainable development in Serbia.

Database of Planned Land Use (PLU)

The initial national database of planned land use (PLU) data has been developed based on valid spatial plans from 18 pilot cities and municipalities, in accordance with the INSPIRE Specification for the Land Use theme and the relevant national rulebook [21]. The database includes five primary land use categories -construction, agricultural, forestry, water, and other purposes (e.g. mining, recultivation and remediation, and defence)- identifiable at the level of individual cadastral parcels (Fig. 11).

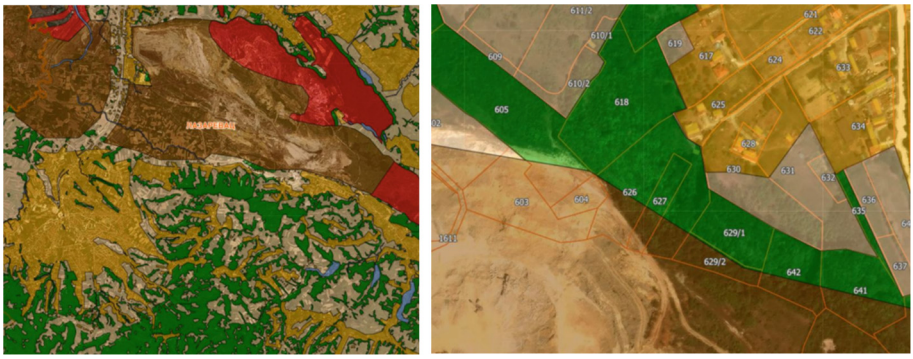


Fig. 11. Overview of data stored within the national Planned Land Use (PLU) database

Although the PLU database currently does not provide more detailed classifications within these general land use categories (e.g. residential, public, commercial, and others), when combined with data from the three Catalogues, it provides users with a sufficiently comprehensive understanding of the development potential of each cadastral parcel.

4.3 Towards Geospatial Platform for Enhancing Cross-Fertilisation Between Spatial Planning and Territorial Cohesion Policies

Despite their complementary roles, the degree of cross-fertilisation between spatial planning and national territorial cohesion policies (e.g. agriculture, environment, natural and cultural heritage protection, transport, telecommunications) remains limited and inconsistent. This is largely due to shifting political priorities, a lack of transparency, misaligned timeframes and, most importantly, insufficient integration -and in some cases absence- of sectoral geospatial data.

According to Serbian regulations, the preparation of spatial and urban plans includes, as a mandatory step, the collection of sectoral requirements (policies, strategies and plans), along with related geospatial and other data from relevant sectoral authorities (institutions and agencies) and public utility providers [22]. These inputs form the

basis for comprehensive thematic (spatial and non-spatial) analyses and the development of thematic planning solutions, followed by a cross-thematic synthesis that defines development priorities for the planning area.



Fig. 12. The SmartParcel system support integration of geospatial data from different sectors, thus efficiently and effectively enhancing their cross-fertilisation for balanced and cohesive territorial development: integrated view of disaster risks (cadastre of landslides and unstable slopes) [33] and planned land use data within city of Kragujevac area (top); integrated view of protected natural and cultural heritage data on the cadastral parcel level downtown city of Belgrade (bottom)

However, this process is often time-consuming and is conducted with limited use of geospatial technologies, which could otherwise enhance efficiency and support transparency, improve the quality of spatial analysis, and enable data integration for better insights. In this context, the SmartParcel system -conceived as a comprehensive NSDI-based geospatial platform- has the potential to provide a central, reliable and permanent collaborative environment. It enables a transparent and shared workspace for spatial

planning stakeholders, ensuring continuous access to and visibility of spatial and sectoral plans, as well as their standardised datasets on planned and ongoing developments across all governance levels.

Although the uploading of sectoral datasets and the population of the three Catalogues are still ongoing, the SmartParcel system already demonstrates functionalities that can facilitate and enhance cross-fertilisation between territorial development data and policies. This is particularly relevant for achieving cohesive, sustainable, resilient and balanced territorial development in Serbia.

Even though the currently available datasets are not yet fully integrated into ongoing plan preparation processes, the platform enables their dynamic integration and visualisation. This supports future cross-fertilisation, for example, between disaster risk reduction strategies and spatial development strategies, or between nature and cultural heritage protection policies and land use planning at the cadastral parcel level (Fig. 12).

In addition to strengthening cross-fertilisation between spatial planning and national territorial cohesion policies, the SmartParcel platform also provides a foundation for future cross-fertilisation and integration with the European CP. Its functionalities align with the approaches and recommendations outlined in Sect. 2 of this paper, highlighting its potential to support more effective, data-driven and place-based policy coordination.

5 Conclusions and Next Steps

By providing integrated geospatial data and related information, the SmartParcel system represents a significant step towards more transparent, efficient, and legally secure land management in Serbia. Furthermore, it supports sustainable, territorially cohesive, and risk-informed decision-making and policy design for land and urban development, in line with modern societal needs and grounded in established global and European geospatial standards and best practices.

As demonstrated in other European countries, the digitalisation of national spatial planning system and practice can substantially improve efficiency, transparency and the quality of planning data and development decisions. In this context, geospatial platform such as SmartParcel provides not only technological advancement but also essential support for cross-fertilisation and the integrated coordination of strategic territorial cohesion policies and instruments, thereby contributing to balanced territorial, social and economic development.

Looking ahead, and in support of improved understanding of territorial dynamics and development opportunities in Serbia, the SmartParcel system will be further developed as part of the continuation of the SmartSDI project. In addition to the anticipated changes in the upcoming Cohesion Policy programming period 2027–2034, future development of the SmartParcel system will also take into account expected updates related to the simplification of the INSPIRE Directive and the adoption of regulations under the GreenData4All initiative [34], within the framework of the European data strategy.

In addition, further research and development of the SmartParcel system (towards the future development of the ePlan concept) will focus on enhancing its positive social impact, particularly in addressing existing gaps between urban and rural areas and promoting the participation of non-expert users in spatial and urban planning processes in

Serbia. This will build on the system's inherent openness, enabled by its technological design, which provides location-independent access and communication advantages.

References

1. Anttiroiko, A.V.: Digital urban planning platforms: the interplay of digital and local embeddedness in urban planning. Int. J. E-Plan. Res. **10**(3), 35–49 (2021). https://doi.org/10.4018/IJEPR.20210701.oa3
2. Cotella, G.: Editorial: EU cohesion policy and domestic territorial governance. What chances for cross-fertilization? Europa XXI **35**, 5–20 (2018). https://doi.org/10.7163/Eu21.2018.35.1
3. EC STRAT-Board. https://territorial.ec.europa.eu/sustainable-development/strategies/strat-board/?lng=en. Accessed 6 Mar 2026
4. EC: The future of cohesion policy. https://ec.europa.eu/regional_policy/policy/how/future-cohesion-policy_en. Accessed 3 Mar 2026
5. EEA: Net land take in cities and commuting zones in Europe. https://www.eea.europa.eu/en/analysis/indicators/net-land-take-in-cities. Accessed 3 Mar 2026
6. ESPON: COMPASS - Comparative Analysis of Territorial Governance and Spatial Planning Systems in Europe (2018). https://archive.espon.eu/sites/default/files/attachments/1.%20COMPASS_Final_Report.pdf
7. ESPON: Cross-fertilisation of cohesion policy and spatial planning, Policy Brief (2021). https://archive.espon.eu/sites/default/files/attachments/ESPON%20Policy%20Brief_Cross-fertilisation%20of%20cohesion%20policy%20and%20spatial%20planning.pdf. Accessed 3 Mar 2026
8. ESPON: No net land take trajectories Policies and practices across Europe, Policy Brief (2024). https://www.espon.eu/sites/default/files/2024-12/no-net-land-take-trajectories_policy-brief.pdf. Accessed 3 Mar 2026
9. ESPON DIGIPLAN: Evaluating spatial planning practices with digital plan data. Final report (2021). https://www.espon.eu/digiplan. Accessed 3 Mar 2026
10. EU INTEGRA programme. https://euintegra.org.rs/en
11. Ferry, M., et al.: Territorial Instruments of National Policies as a Tool for Territorially Sensitive Policies (2025). https://www.espon.eu/sites/default/files/2026-01/senpo_final_report.pdf. Accessed 3 Mar 2026
12. GIZ: EU support for the Republic of Serbia to effectively implement Cohesion policy. https://www.giz.de/en/projects/eu-support-republic-serbia-effectively-implement-cohesion-policy. Accessed 3 Mar 2026
13. Grilc, M., et al.: Систем Паметна Парцела – Темељ дигиталног простора, Научно-стручна конференција са међународним учешћем "Урбанизам и одрживи развој" (2025). https://doi.org/10.46793/Urbanizam25.129G
14. Grilc, M., Gruevski, G.: Систем Паметна Парцела – Темељ дигиталног простора (2025). https://pp.geosrbija.rs/pp-home-ui/kb-single/527
15. Janowski, T., Estevez, E., Baguma, R.: Platform governance for sustainable development: reshaping citizen-administration relationships in the digital age. Gov. Inf. Q. **35**(4), S1–S16 (2018). https://doi.org/10.1016/j.giq.2018.09.002. Supplement
16. Lacoere, et al.: National introduction of no net land take: a comparative study of five pioneering countries seeking to limit their land consumption. Town Plan. Rev. **96**, 347–371 (2025). https://doi.org/10.3828/tpr.2024.44
17. Maier, K., Kabrhel, J., Dabrowski, M.: Cross-fertilisation between spatial planning and territorial cohesion: lessons from the Czech Republic. Plan. Pract. Res. **40**(1), 143–163 (2025). https://doi.org/10.1080/02697459.2024.2416352

18. Ministry of construction, transport and infrastructure: Стратегија одрживог урбаног развоја Републике Србије до 2030. Године (2019). https://pravno-informacioni-sistem.rs/eli/rep/sgrs/vlada/strategija/2019/47/1/reg
19. Ministry of construction, transport and infrastructure: Јавни увид у нацрт Просторног плана Републике Србије од 2021. до 2035. године и извештај о Стратешкој процени утицаја Просторног плана Републике Србије од 2021. до 2035. године на животну средину (2021). https://www.mgsi.gov.rs/cir/dokumenti/javni-uvid-u-nacrt-prostornog-plana-republike-srbije-od-2021-do-2035-godine-i-izveshtaj-o
20. Ministry of construction, transport and infrastructure. LIID Project. https://liid.mgsi.gov.rs/en/liid-project/. Accessed 3 Mar 2026
21. Official Gazette of the Republic of Serbia: Rulebook on Content and Method of Keeping and Maintaining the Central Register of Planning Documents, Information System on the State of the Environment and Local Information System, and on the Digital Format for Submission of Planning Document (2015). https://pravno-informacioni-sistem.rs/eli/rep/sgrs/ministarstva/pravilnik/2015/33/3/reg
22. Official Gazette of the Republic of Serbia: Rulebook on the Content, Method, and Procedure for Preparation of Spatial and Urban Planning Documents (2025). https://pravno-informacioni-sistem.rs/eli/rep/sgrs/ministarstva/pravilnik/2019/32/2/reg
23. Republic Geodetic Authority: Strategic action plan for the national spatial data infrastructure of the Republic of Serbia (2021). https://en.rgz.gov.rs/content/pages/english/Structure/NSDI%20Strategic%20Actions%20Plan-4.pdf. Accessed 12 Mar 2026
24. Republic Geodetic Authority: Analysis of Results Acquired through the On-line Questionnaire on the Interest of the Municipalities and Cities to Participate in the SmartSDI Project Implementation (2024). https://pp.geosrbija.rs/pp-home-ui/kb-single/756
25. UNGGIM: Framework for Effective Land Administration, August 2021, New York, United States. https://ggim.un.org/meetings/GGIM-committee/10th-Session/documents/E-C.20-2020-29-Add_2-Framework-for-Effective-Land-Administration.pdf. Accessed 3 Mar 2026
26. UNGGIM: United Nations Integrated Geospatial Information Framework (UN-IGIF). https://www.un.org/globalgeospatial/en/united-nations-integrated-geospatial-information-framework-un-igif
27. World Bank: Green, Livable, and Resilient Cities, Serbia: Strengthening Sustainable and Resilient Urban Development (English). Washington, D.C.: World Bank Group (2023). http://documents.worldbank.org/curated/en/099102023122539375. Accessed 12 Mar 2026
28. World Banka: GREEN, LIVABLE, RESILIENT CITIES IN SERBIA: Comparative analysis of 10 cities. (English). World Bank Group, Washington, D.C. (2023). https://thedocs.worldbank.org/en/doc/41e8d750ee019f0d209b83be6703670c-0080012024/original/Serbia-Comparative-Analysis-of-Ten-Cities.pdf
29. World Banka: Sustainable Cities Serbia: Unlocking the transformational potential of cities for the green transition (English). World Bank Group, Washington, D.C. (2023). http://documents.worldbank.org/curated/en/099102023122572634
30. Živković, L.: Monitoring urban development: national register of investment locations as a tool for sustainable urban land use management in Serbia. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. LNCS, vol. 12251. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58808-3_59
31. Živković, L.: Project e-space: building a digital platform for spatial and urban planning and development in Serbia. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2021. ICCSA 2021. LNCS, vol. 12953. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86976-2_34

32. Živković, L.: Project Smart SDI: concept for improving NSDI for sustainable development in Serbia. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2023 Workshops. ICCSA 2023. LNCS, vol. 14105. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37108-0_35
33. Živković, L.: National spatial data infrastructure (NSDI) for resilient territorial development: building a national disaster risk register (DRR) for Serbia. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., Rocha, C.A.M.A., Faginas Lago, M.N. (eds.) Computational Science and Its Applications – ICCSA 2024 Workshops. ICCSA 2024. Lecture Notes in Computer Science, vol. 14817. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-65238-7_12
34. EC – Directorate-General for Environment: Simplification of administrative burdens in environmental legislation (2025). https://environment.ec.europa.eu/publications/simplification-administrative-burdens-environmental-legislation_en
35. Official Gazette of the Republic of Serbia: Law on Planning and Construction (2025). https://pravno-informacioni-sistem.rs/eli/rep/sgrs/skupstina/zakon/2009/72/11/reg

Author Index

A
Abreu, Francisco 115
Acebrón, Juan 542
Agra, Agostinho 283
Alagia, Michele 436
Almeida, Nuno 115
Almeida, R. De 301
Alvelos, Filipe 283, 314
Alvelos, Helena 283
Alves, Cláudio 266
Andreatta, Karina Assini 351
Aoyama, Shigeyoshi 492
Asahiro, Yuichi 231

B
Bairrada, Cristela 167
Balletto, Ginevra 577, 610
Balucani, Nadia 425, 448, 461
Baptista, José 215
Boeres, Maria Claudia Silva 351

C
Caroni, Chrys 3
Cavalli, Simonetta 436
Ceccarelli, Cecilia 461
Cerveira, Adelaide 215
Codella, Claudio 461
Correia, Aldina 87
Costa, Lino 182

D
Daisaka, Hiroshi 477
de Doncker, Elise 477
de Moraes, Renato Elias Nunes 351
de Oliveira, Sanderson L. Gonzaga 199
de Sousa, Alexandre Valente Conceição Pereira 563
Deus, Rui Gonçalves de 51
Di Genova, Gabriella 461
Dias, Joana 167
Doğruluk, Ertuğrul 563
dos Santos, Francisco 182

E
Enomoto, Takeshi 525

F
Falcinelli, Stefano 436
Farsimadan, Eslam 387, 406
Feifel, Raimund 436
Fernandes, Ana 149
Ferreira, Maria J. 87
Filomena Teodoro, M. 99

G
Gatta, Mário 69
Ghiani, Emilio 577
Gouiaa-Mtibaa, Amal 20

H
Hiyama, Emiko 492

I
Ibrahim, Asim 371
Ibrar, Kainat 371
Ishikawa, Tadashi 477
Isola, Federica 577

K
Kanamori, Issaku 492
Kato, Kiyoshi 477
Kouya, Tomonori 508
Kratz, Frédéric 20

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16760, pp. 659–660, 2027.
https://doi.org/10.1007/978-3-032-30521-3

L
Ladu, Mara 577, 610
Lagarias, Apostolos 622
Lima, Maria Coelho 314
Lima, Vanda 87
Lucena, Rui 115
Lukačínová, Michaela 592

M
Mancini, Luca 425, 448, 461
Marques, Francisco 283
Matsufuru, Hideo 492
Mendes, Rodrigo João Bastos 563
Mieri, Lucas E. 199
Miguel, Belchior 51
Milesi, Alessandra 577
Mital', Ondrej 592
Miyano, Eiji 231
Monteiro, Ana M. 149
Monteiro, José 542
Moradi, Leila 387, 406

N
Nakamura, Shunta 231
Nakasato, Naohito 477
Natário, Isabel 51

O
Oliveira, Daniel 542
Oliveira, Teresa A. 132
Oliveira, Teresa 99

P
Palmieri, Francesco 371, 387, 406
Pannacci, Giacomo 448
Papazoglou, Angeliki 622
Parriani, Marco 425, 436, 448
Pascoal, Rui 149, 332
Pereira, P. Sampaio 69
Pierri, Francesca 3
Pimenta, Pedro Correia Cravo 563
Pimentel, Carina 314
Pinto, Telmo 266
Pirani, Fernando 436
Podio, Linda 461

R
Regis, Rommel G. 247
Ribeiro, Diogo 87
Richter, Robert 436
Rocha, Humberto 332
Rosi, Marzio 425, 448, 461
Ručinská, Silvia 592

S
Santos, Patrícia 167
Santos, Renato 115
Sendas, Joana 332
Shimizu, Noritaka 492
Silva, Anaís Veloso 266
Silva, Paulo 215
Simões, Paula 51, 115
Skouteris, Dimitrios 448, 461
Stranges, Stefano 436
Stratigea, Anastasia 622
Sushchenko, Olha 37

T
Teixeira, A. 301
Teodoro, M. Filomena 69
Teodoro, Maria Filomena 132
Trovati, Marcello 387, 406
Tsunoda, Yusuke 492

V
Vanuzzo, Gianmarco 448
Varela, Leonilde 182
Vecchiocattivi, Franco 436
Vrignat, Pascal 20

W
Wardhana, Indrawata 20

X
Xambre, Ana Raquel 283

Y
Yuasa, Fukuko 477

Z
Živković, Ljiljana 641