

Machine Learning Based Approach to Assess Territorial Marginality

Simone Corrado(✉) and Francesco Scorza

Laboratory of Urban and Regional System Engineering (LISUT), School of Engineering, University of Basilicata, Potenza, Italy
simone.corrado@studenti.unibas.it

Abstract. The territorial cohesion is one of the primary objectives for the European Union and it affects economic recovery pushing the role of Public Administration in promoting territorial development actions. The National Strategy for Inner Areas (SNAI) is a public policy promoting endogenous development processes in marginal territories with low settlement density. Specific contexts where rules and standards defined for the organization of large metropolitan aggregates lose their effectiveness whose identification represents a critical stage for policy efficacy and the actual map of SNAI target areas appears to be the results of a weak and simplified analytical approach. These considerations are the origin of the research question that underlies this work: identify the typical characteristics of Basilicata's marginal areas through machine learning techniques and, subsequently, reclassify the national territory using the trained model. However, outlining the boundary of this territories is only a preliminary task. The following step is the identification of the dynamics within the different territorial sub-systems that make up the inner peripheries. This paper presents the results of the local model-agnostic method for interpreting the obtained results. It emerges, thought cooperative game theory by Shapley values, the need to refine analytical methods that are sensitive to the measurement of the different context conditions. Future perspectives of the research regard the extensive deepening of the application on the basis of wider datasets able to make explicit spatial components of the distribution of the observed phenomena.

Keywords: Territorial cohesion · Inner periphery · Interpretable machine learning

1 Introduction

The territorial cohesion is one of the primary objectives for the European Union [1, 2]. Rural areas host almost a third of the EU population, and in the Italian national context about a quarter of the entire population lives in marginal areas: the inner peripheries [3, 4]. The Italian territory is organized in a multiplicity of minor centers connected to main metropolitan poles [5]. This spatial arrangement brings out the extraordinary variety of inner areas at all levels i.e., geographic, economic, social, cultural, eco-systemic. The identification of the differences between local systems of marginal areas is the first step

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022, LNCS 13376, pp. 292–302, 2022.
https://doi.org/10.1007/978-3-031-10450-3_25

in recognizing the intrinsic complexity of these territories. Hence arise the needs for integrated spatial planning based on sustainable development, a substantial and particularly critical exercise in the management of so-called weak demand territories, i.e., with low settlement density, where rules and standards defined for the organization of large metropolitan aggregates lose their effectiveness [6]. In fact, Italian inner areas have been facing process of marginalization since the 1950s. Intense phenomena of deantrophization have led to a substantial decrease in population and progressive demographic aging, the reduction of the active population with a consequent reduction in employment and under-use of local territorial resources [7]. Therefore, inner peripheries are considered a national issue, because of their development potential, as well as their community costs due to their social condition and the unequal access to basic services compared to the more developed centers [8].

The national strategy for inner areas seeks to reverse these trends, through a gradual, integrated and monitored approach and by evaluating land governance and development strategies [9]. Such innovative policy aims to improve the quality of life of residents, contributing to national economic recovery and the achievement of the European Union's territorial cohesion objectives [10]. In organizing such complex policy framework, the primary issue is the identification of the boundaries of the target areas. SNAI reduces this issue to an operative simplification based on a limited set of indices at national level.

In this paper a broader analytical procedure is proposed for the identification of inner areas according to a robust spatial appraisal including multivariate datasets. We base such analysis on the application of machine learning techniques to a database including socio-economic indicators and variables descriptive of the morphological characteristics linked to spatial measures of accessibility. This approach tends to improve the selection of target areas for the policies promoted by the SNAI by bringing out different functional specializations on the basis of which endogenous development policies can be better structured. In particular, the analytical approach takes into account, for the definition of the reference labels, the characteristics of the marginal territories identified in the Basilicata Region. The working hypothesis it's to consider Basilicata as a benchmark role for the model. The results obtained propose a new map of the inner areas that is scarcely comparable with that of the SNAI. It opens up critical considerations on the potential impact of the development support policy and confirms the need to refine analytical methods that are sensitive to the measurement of the different context conditions. The conclusions highlight how the mapping of territorial marginality obtained through the proposed analytical approach is, in fact, conditioned by specific structural variables such as "resident population", characteristics linked to "ageing", and the "low productive specialization" of municipalities. They confirm the need to deepen the application on the basis of more extensive datasets able to make explicit spatial components of the distribution of the observed phenomena.

2 Materials and Method

The previous section already introduced the growing importance of data in the progress of spatial development strategies for the future [11, 12]. To better explain what big-Data is, it is necessary to specify what is meant as data. The DIKW method is introduces as

a hierarchical framework to get insight and find pattern into data [13]. Data are closely related to places, are potentially reproducible, are part of larger information structures, and can be processed and stored in digital formats [14]. Raw-data itself are neutral to the human mind, they become information when they are extracted, processed and used for certain purposes. The framework of information often consisting of multiple data from the same and different types of domains becomes knowledge when they are interpreted through tools, applications, methods, cross-references, indicators etc. [15]. Knowledge, finally, gains the value of awareness -in the case of Open Data it can certainly be defined as collective, in the broadest sense of common benefit- when it is able to perform well the complexity of the real world [16]. The issues in big-data management are well known and concern volume, velocity and variety without ignoring two additional dimensions: variability and authenticity of the information [17].

The SNAI database was chosen as basic dataset for this research. This is an informative infrastructure of accessible and interoperable data certified by the national agency of territorial cohesion in which the minimum statistical unit is the municipality. It contains variables and indicators for individual municipality related to the scopes of the SNAI Strategy. The thematic areas analyzed are: social, demographic, economic situation and consistency of essential services. The data are extremely heterogeneous both by topic and year of release. Some variables are collected from 1971 to 2011 on a 10-year basis for statistical trend setting others cover only the year 2011. The following Table 1 gives a summary of the characteristics of SNAI database.

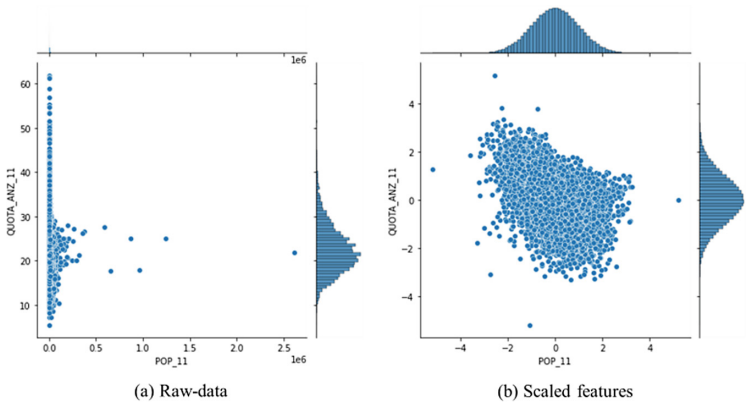
Table 1. Summary of SNAI database

Context	Indicator	Variable	Total
Demography	9	13	22
Economy	10	19	29
Morphology	6	6	12
Health	3	5	8
School	3	51	54
Transport	1	5	6
Digital divide	2	-	2
Total	34	99	133

Supervised Machine Learning (ML) techniques emerge to be the most capable at handling such informational heterogeneity. As known, ML is a set of methods that computers use to make and improve predictions or behaviors based on data. Specifically, supervised learning allows finding patterns between data and labels that can be expressed mathematically as functions [18]. The model will then learn from train data by identifying relationships in feature space - refers to the n-dimensions where variables are distributed - over the homogeneous reference group given by the labels. Then the ML model will be used to make predictions about the new data not assessed during training phase. For this

property, the ML model will be used to make predictions about the unseen municipality data. Python programming language was preferred because it integrates many ML and data visualization libraries. In particular, the Numpy and Pandas libraries were used as data analysis and manipulation tools, scikit-learn for data engineering and ML models, and the seaborn library for visualize, explore and understand the data [19–21].

To achieve the aim of the research, it is necessary to deconstruct the analysis problem into distinct moments. Firstly, more than one million data from the 8092 Italian municipalities were analyzed to seek for possible statistical mismatches and to assess their frequency distribution. The in-depth analysis identified the transformation most akin to the distribution of the data and the best techniques to impute the missing values. Commonly, ML algorithms don't perform well if feature have very different scales, -in SNAI database data varies widely, some indicators are percentages, while other demographic variables ranging from 1e3 to 1e6 and so on-, thus a feature scaling is needed [22]. In fact, most of the features have a lognormal distribution with metropolitan cities set as outliers for many variables. After this validation check, data engineering process was performed. In particular, data have been transformed for stabilizing variance and minimizing skewness with a quintile transform feature-wise to make distribution more Gaussian-like (see Fig. 1).



(a) Raw-data (b) Scaled features

Fig. 1. Comparison between raw (a) and transformed (b) data distribution of residential population and demographic aging index.

Labels of the Basilicata municipalities are assumed from a previous work that investigated the hierarchy of policentricity ranks according to Christaller's Central place theory and PCA statistics performed on in deep analysis of a peculiar set of variables describing the services supply and the rarity index for services categories as a way to detect territorial marginality [23, 24].On the basis of the frequency of each service category, an aggregate assessment of the rank of each Basilicata municipality was drawn up in 4 classes (R1- main urban center, R2, R3, R4 – inner periphery). This is not a simple ranking of the cities, but a hierarchical dependence, in the sense that those of higher rank also encompass in their sphere of superior services areas gravitating for common

services on cities of lower rank. Thus, in this work all municipalities classified as the lowest rank are considered inner peripheries.

The ML classification algorithm used is the "decision-tree" (see Fig. 2). It allows a significant interpretation of the territorial marginality since it is more closely mirror human decision-making. In facts, tree-based models split the data multiple times in order to obtain subsets which minimize the Gini index or the entropy of the system. The algorithm hence tends to search within the dataset for those predictors that best split the sample into pure subsets, that is, data that all fall into the same class. The model therefore predicts that each new observation belongs to the class most commonly found in training observations in the rank to which it belongs. The algorithm continues to split the data recursively until the Gini index drop to zero. This process produces a tree that resulting in a perfectly fit on training set, but is likely to over-fit the data, leading to a poor generalization [22]. A better strategy is to grow a smaller tree with fewer data splits in order to have lower variance and a better interpretation at the cost of purity leaf split. Thus, a tree pruning is set on the fourth leaf. However, after an initial attempt with a single decision-tree model, a tree bagging was carried out through the use of random forest algorithm to overcome the over-fitting issues and substantially improve the predictive performance.

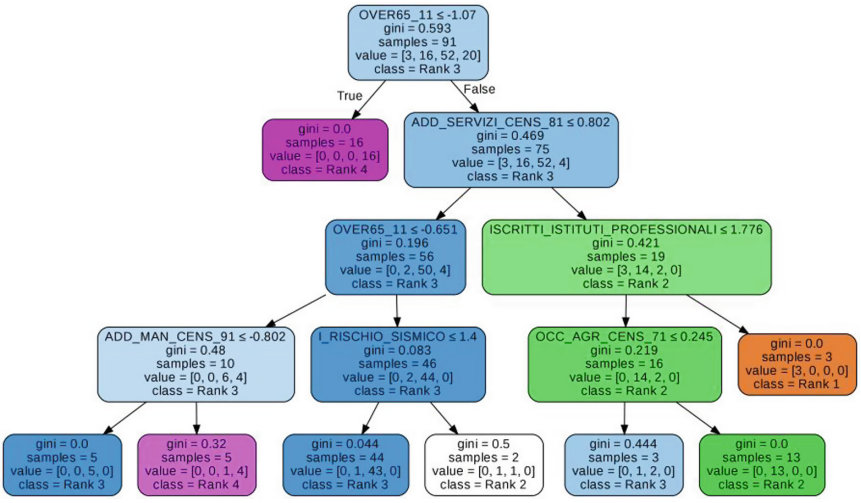
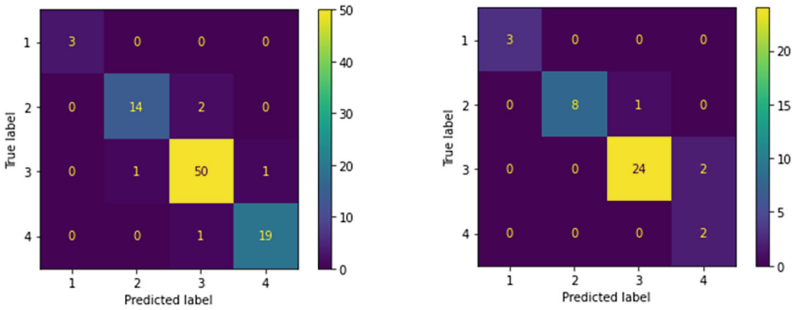


Fig. 2. Single decision-tree in which the different colors specify the rank classification.

3 ML Interpretability and Results

Trained ML models are designed to make the most accurate predictions from data. The accuracy of the model allows to evaluate its effectiveness and avoid the over-fitting issue. There are several metrics to assess multi-label classification score. The simplest one is accuracy score. This function computes general model accuracy com-paring the set of

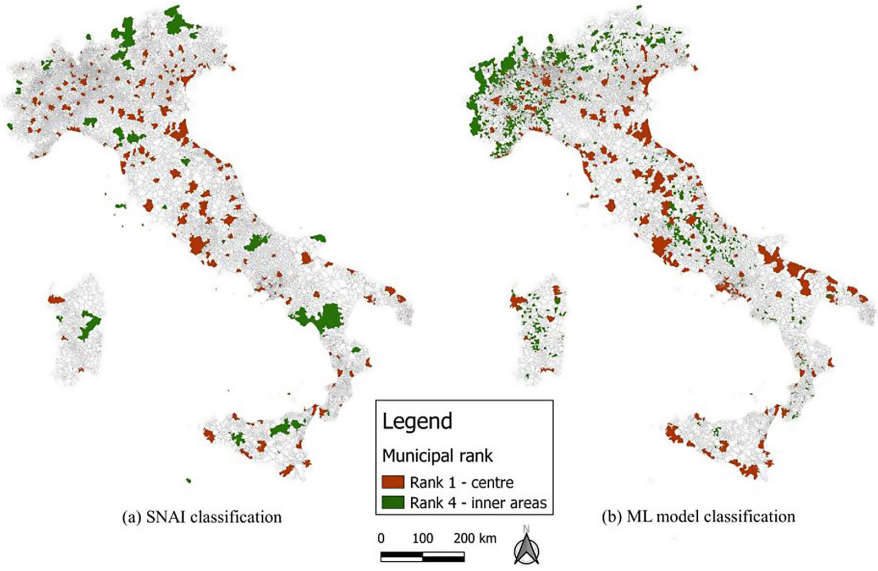
labels predicted for a sample against the same subset true label. The output model has a score on the train subset of 0.94 instead on the test subset equal to 0.87. Also, a confusion matrix can be displayed. This metric makes it possible to investigate in more depth on which group the model fails (see Fig. 3).



(a) Train set (b) Test set

Fig. 3. Confusion matrix on train (a) and test (b) split database.

The overall accuracy of the classification model is considered adequate, and the confusions matrix analysis allows to estimate negligible losses over rank 4. Once the model was trained and tested, the classification of municipal ranks was transferred to the Italian municipalities (See Fig. 4).



(a) SNAI classification (b) ML model classification

Fig. 4. Classification comparison between SNAI (a) and trained ML algorithm (b).

The classification of the inner areas obtained with the trained ML model is scarcely comparable with the SNAI. The resulting classification is much more fragmented. This is mainly due to the hypothesis to consider as labels for inner areas those municipalities that in the Basilicata hierarchy was the weakest and rarest ones. In other words, those territory presenting the highest development demand on the basis of the regional conditions. Also, the pole cities present a different geography. Metropolitan areas are identifiable in the results and their scope is even wider. This depends also on the training set where Potenza and Matera (effectively two small town) are classified in the highest rank as urban poles. Moreover, the greatest differences are found in the west Alpine region included in the inner peripheries by the ML model while in the SNAI classification it was not considered an inner area for such extend. What also emerges is the fragmented classification throughout the whole Apennine bone. This also depends on the characteristics of the training subset used for ML. Eventually additional appraisal of morphological features and more accurate accessibility analysis may improve such result.

A relevant stage in ML classification is the interpretation of the model [25]. It means going into the specific informative contribution of each variable and understanding its relevance within it and overcome the concern of black-box model. In detail, interpretable ML refers to methods and models that make the behavior and predictions of machine learning systems more understandable to humans [26–28]. As previously mention, decision tree models are very interpretable because they are similar to human thinking and decision-making. In fact, a single decision tree can be displayed to understand how it's work and why it used some feature rather than others to split the subsets. Instead, the Random-forest algorithm improve prediction accuracy at the expense of interpretability but this is a necessary trade-off for having a more flexible approach. Through cooperative game theory approach also this methodology becomes more intuitive assuming that each feature value of the instance is a "player" in a game where the correct prediction is the payout [29]. In fact, Shapley value is a system for distributing a reward obtained by the feature among its coalition subset, and its purpose is to distribute that reward proportionally to the contribution each player makes to the coalition. The distribution of Shapley values for each feature allows to investigate the ML model through a local model-agnostic method. Thus, the impacts of features on the model's choices are estimated with the SHAP package. The following summary plot represent the feature relative importance and also their impact on model output (See Fig. 5).

Depicting to the finding, assumptions regarding inner peripheries are validated. Socio-economic rather than morphological characteristics defined the inner peripheries. These areas present a general situation of structural demographic weakness with a trend that continues to worsen due to inertia. In fact, the impact on model choices is significant. This is reflected also in a shortage of employees in the economic sectors, especially in the tertiary sector, with significant percentage changes in employment from 1971 to 2001. A proxy is also the high seismic risk territories that enhance the isolation of these areas.

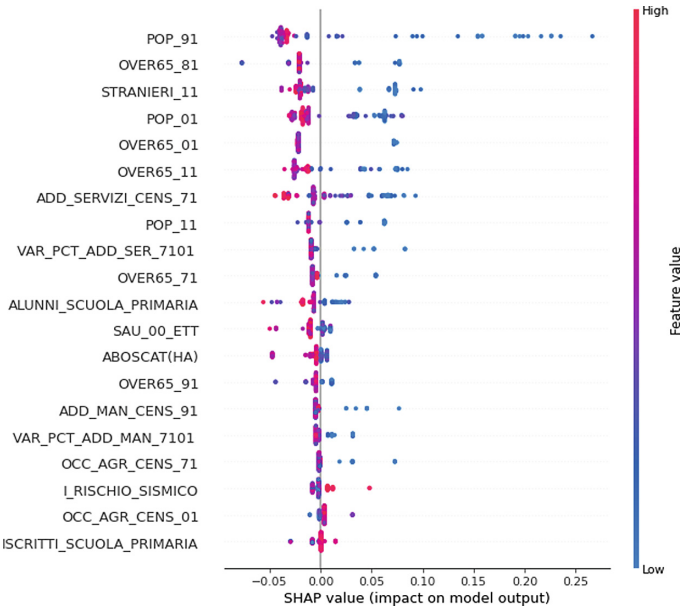


Fig. 5. Summary plot of Shapley values for inner areas

4 Conclusions

ML and model interpretation have proved to be two effective tools supporting the territorial research for the identification of inner areas. Among results, detecting features relevance in the model represents a first driver to better orient the strategic actions: i.e., if ageing population represents a key component in inner areas classifications, the policy should prioritize actions oriented to support the needs for those inhabitants and to adopt measure to reduce the relative weight of this population components in favor of the others ones. This analytic approach is meaningful also in validating existing assumptions about inner peripheries, grasp and anticipate changes that might set-in and design place-based development policies, put in practice effective territorial monitoring tools.

Measuring the effects of the SNAI strategy on demographics therefore becomes a key-factor in assessing the impact of local development policies and projects [30]. However, a study has already shown that the SNAI database requires revision both in the accuracy of some municipal scale indicators and in the consistency of some features [31]. In fact, the database is unbalanced towards demographic feature rather than economic and especially geo-morphological variables. Such unbalanced classes could affect the deciphering of patterns within the data [32]. Reference studies to be considered in order to extend the informative bases regards: accessibility and mobility management [33, 34] in urban and territorial scale, impact of regional development policies funded by EU [35, 36], territorial fragmentation and anthropic pressure on relevant ecosystem values [37–40], energy transition and climate proof municipalities [41–44]. Future research perspectives are: revisiting the database by identifying additional Open Government dataset that have been activated since 2011, and integrating trends and indices deemed

useful to the objectives of the National strategy. Furthermore, spatial inequality and segregation will be investigated to identify also the relationship between the inner area for implementing effective community management policies.

References

1. Dax, T., Copus, A.: Towards vibrant, inclusive and sustainable rural regions: orientations for a future rural cohesion policy. Eur. Struct. Invest. Funds J. **6**, 198–210 (2018)
2. European Commission: ESDP European Spatial Development Perspective: Towards Balanced and Sustainable of the European Union (1999)
3. Matthews, A.: Rural Development in the European Union: Issues and Objectives, pp. 1–14. Citeseer (2007)
4. Eurostat: Statistics on regional typologies in the EU - Statistics Explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Statistics_on_rural_areas_in_the_EU#Source_data_for_figures_and_maps_.28MS_Excel.29. Accessed 24 May 2022
5. Eurostat: Territorial typologies - Statistics Explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Territorial_typologies#Typologies. Accessed 24 May 2022
6. Scorza, F., Saganeiti, L., Pilogallo, A., Murgante, B.: Ghost planning: the inefficiency of energy sector policies in a low population density region. Arch. di Stud. Urbani e Reg. 34–55 (2020). https://doi.org/10.3280/ASUR2020-127-S1003
7. Dax, T., Fischer, M.: An alternative policy approach to rural development in regions facing population decline. Eur. Plan. Stud. **26**, 297–315 (2018). https://doi.org/10.1080/09654313.2017.1361596
8. Balzan, M.V., Caruana, J., Zammit, A.: Assessing the capacity and flow of ecosystem services in multifunctional landscapes: evidence of a rural-urban gradient in a Mediterranean small island state. Land Use Policy **75**, 711–725 (2018). https://doi.org/10.1016/j.landusepol.2017.08.025
9. Carrosio, G.: A place-based perspective for welfare recalibration in the Italian inner peripheries: the case of the Italian strategy for inner areas. Sociol. E Polit. Soc. 50–64 (2016). https://doi.org/10.3280/sp2016-003004
10. Knickel, K., et al.: Between aspirations and reality: making farming, food systems and rural areas more resilient, sustainable and equitable. J. Rural Stud. **59**, 197–210 (2018). https://doi.org/10.1016/j.jrurstud.2017.04.012
11. Athey, S.: Beyond prediction: using big data for policy problems. Science. **355**, 483–485 (2017). https://doi.org/10.1126/SCIENCE.AAL4321
12. Höchtl, J., Parycek, P., Schöllhammer, R.: Big data in the policy cycle: policy decision making in the digital era. J. Organ. Comput. Electron. Commer. **26**, 147–169 (2016). https://doi.org/10.1080/10919392.2015.1125187
13. Rowley, J.: The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. **33**, 163–180 (2007). https://doi.org/10.1177/0165551506070706
14. Murgante, B., Borruso, G., Lapucci, A.: Geocomputation and Urban Planning. Springer, Heidelberg (2009).https://doi.org/10.1007/978-3-540-89930-3_1
15. Zins, C.: Conceptual approaches for defining data, information, and knowledge. J. Am. Soc. Inf. Sci. Technol. **58**, 479–493 (2007). https://doi.org/10.1002/asi.20508
16. Las Casas, G., Murgante, B., Scorza, F.: Regional local development strategies benefiting from open data and open tools and an outlook on the renewable energy sources contribution. In: Papa, R., Fistola, R. (eds.) Smart Energy in the Smart City. GET, pp. 275–290. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-31157-9_14

17. Laney, D.: others: 3D data management: controlling data volume, velocity and variety. META Gr. Res. note. **6**, 1 (2001)
18. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
19. McKinney, W.: Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, pp. 56–61 (2010). https://doi.org/10.25080/majora-92bf1922-00a
20. Harris, C.R., et al.: Array programming with NumPy. Nature **585**, 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2
21. Waskom, M.: seaborn: statistical data visualization. J. Open Source Softw. **6**, 3021 (2021). https://doi.org/10.21105/joss.03021
22. Hastie, T., Tibshirani, R., James, G., Witten, D.: An Introduction to Statistical Learning (2nd ed.). Springer Texts, vol. 102, 618. Springer, New York (2021). https://doi.org/10.1007/978-1-0716-1418-1
23. Getis, A., Getis, J.: Christaller's central place theory. J. Geog. **65**, 220–226 (1966)
24. Curatella, L., Scorza, F.: Una Valutazione della struttura policentrica dell'insediamento nella Regione Basilicata. LaborEst. **20**, 37–42 (2020)
25. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning (2017). https://doi.org/10.48550/arxiv.1702.08608
26. Molnar, C.: Interpretable Machine Learning. A Guide for Making Black Box Models Explainable (2019)
27. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
28. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. U.S.A. **116**, 22071–22080 (2019). https://doi.org/10.1073/PNAS.1900654116/SUPPL_FILE/PNAS.1900654116.SAPP.PDF
29. Shapley, L.S.: A Value for N-Person Games. RAND Corporation, Santa Monica, CA (1952). https://doi.org/10.7249/P0295
30. Las Casas, G., Scorza, F., Murgante, B.: Conflicts and sustainable planning: peculiar instances coming from Val d'agri structural inter-municipal plan. In: Papa, R., Fistola, R., Gargiulo, C. (eds.) Smart Planning: Sustainability and Mobility in the Age of Change. GET, pp. 163–177. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77682-8_10
31. Rossitti, M., Dell'ovo, M., Oppio, A., Torrieri, F.: The italian national strategy for inner areas (SNAI): a critical analysis of the indicator grid. Sustainability. **13**, 6927 (2021). https://doi.org/10.3390/su13126927
32. Géron, A.: Book Review: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd edn. O'Reilly Media, Sebastopol (2019)
33. Scorza, F., Fortunato, G.: Cyclable cities: building feasible scenario through urban space morphology assessment. J. Urban Plan. Dev. **147**, 05021039 (2021). https://doi.org/10.1061/(asce)up.1943-5444.0000713
34. Scorza, F., Fortunato, G., Carbone, R., Murgante, B., Pontrandolfi, P.: Increasing urban walkability through citizens' participation processes. Sustain. **13**, 5835 (2021). https://doi.org/10.3390/su13115835
35. Scorza, F.: Improving EU cohesion policy: the spatial distribution analysis of regional development investments funded by EU structural funds 2007/2013 in Italy. In: Murgante, B., et al. (eds.) Computational Science and Its Applications – ICCSA 2013. LNCS, vol. 7973, pp. 582–593. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39646-5_42
36. Scorza, F., Casas, G.L.: Territorial specialization in attracting local development funds: an assessment procedure based on open data and open tools. In: Murgante, B., et al. (eds.) Computational Science and Its Applications – ICCSA 2014. LNCS, vol. 8580, pp. 750–757. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09129-7_54

37. Curatella, L., Scorza, F.: Polycentrism and insularity metrics for in-land areas. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2020. LNCS, vol. 12255, pp. 253–261. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58820-5_20
38. Scorza, F., Fortunato, G.: Active mobility oriented urban development: a morpho-syntactic scenario for mid-sized town. Eur. Plan. Stud. (2022).https://doi.org/10.1080/09654313.2022.2077094
39. Pilogallo, A., Saganeiti, L., Scorza, F., Las Casas, G.: Tourism attractiveness: main components for a spacial appraisal of major destinations according with ecosystem services approach. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2018. LNCS, vol. 10964, pp. 712–724. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95174-4_54
40. Scorza, F., Pilogallo, A., Saganeiti, L., Murgante, B.: Natura 2000 areas and sites of national interest (SNI): measuring (un)integration between naturalness preservation and environmental remediation policies. Sustainability. **12**, 2928 (2020). https://doi.org/10.3390/su12072928
41. Corrado, S., Giannini, B., Santopietro, L., Oliveto, G., Scorza, F.: Water management and municipal climate adaptation plans: a preliminary assessment for flood risks management at urban scale. In: Gervasi, O., et al. (eds.) Computational Science and Its Applications – ICCSA 2020. LNCS, vol. 12255, pp. 184–192. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58820-5_14
42. Scorza, F., Santopietro, L.: A systemic perspective for the sustainable energy and climate action plan (SECAP). Eur. Plan. Stud. 1–21 (2021). https://doi.org/10.1080/09654313.2021.1954603
43. Santopietro, L., Scorza, F.: The Italian experience of the covenant of mayors: a territorial evaluation. Sustainability. **13**, 1289 (2021). https://doi.org/10.3390/su13031289
44. Santopietro, L., Scorza, F., Murgante, B.: Multiple components in GHG stock of transport sector: technical improvements for SECAP baseline emissions inventory assessment. TeMA J. L. Use Mobil. Environ. **15**, 5–24 (2022). https://doi.org/10.6092/1970-9870/8391