Urban Data-Driven Energy Planning: Insights and Challenges from Italy and Portugal

Carmen Guida[1], Paulo Ricardo Batista[2], and Gerardo Carpentieri[1](✉)

[1] University of Naples Federico II, Naples, Italy
{carmen.guida,Gerardo.carpentieri}@unina.it
[2] University of Aveiro, Aveiro, Portugal

Abstract. Effective urban planning for sustainable and equitable development requires data-driven support for policymakers to enable context-specific actions. Big and Open Data can serve as valuable resources for technicians and policymakers, transforming scientific knowledge into decision-support tools. However, the complexity and lack of standardisation of such data can create significant gaps. In the energy sector, open data issues are even more complex due to multiple stakeholders, privacy concerns, and varying motivations for collecting energy consumption data, such as municipal taxation versus energy management.

This paper investigates these critical issues, which underpin the development of reliable, data-driven methodologies for understanding urban systems and supporting targeted interventions. Through two comparative case studies in Italy and Portugal, we highlight how data structure and accessibility shape the methodological approach and outcomes. In the Italian case study, conducted in Naples, comprehensive energy consumption data were available through the Siatel platform, including detailed records for individual users across the city. While the richness of the dataset enabled high-resolution analysis, the primary challenge lay in accurately geocoding and spatially integrating user-level data within the urban fabric. In contrast, the Portuguese case study relied on more aggregated energy datasets, where the key methodological challenge was disaggregating and translating large-scale information into finer spatial units suitable for meaningful urban analysis.

The comparison demonstrates that even when adopting replicable methodologies and modelling approaches, valid cross-city comparison depends on explicit harmonization of the target variables produced from distinct data infrastructures, because addressing data-specific constraints is essential for producing effective, context-sensitive planning solutions.

Keywords: Energy data · Urban energy planning · Geocoding · Spatial harmonizarion

1 Introduction

The disruptions caused by recent global crises have redefined how planners and researchers conceptualise urban quality of life, revealing with unprecedented clarity the profound impact of residential conditions, such as dwelling quality, neighbourhood

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16764, pp. 501–517, 2027.
https://doi.org/10.1007/978-3-032-30533-6_33

density, access to open space, and the thermal performance of the housing stock, on the health, well-being, and energy vulnerability of urban populations [1–3]. This context makes the goal of building smarter, safer, and healthier cities more important than ever [4]. It also shows what aggregate urban statistics often hide: that there are big differences in how livable different areas are, and that fixing them needs accurate, detailed data at the neighbourhood and building level [5]. The methodological inquiries this paper examines, how to geocode imprecise administrative records and how to disaggregate collective energy statistics into significant spatial units, are not merely technical endeavours. They are necessary steps to identify where residential energy vulnerability is highest, to ensure that retrofit programs reach the people who need them most, and to monitor the long-term effectiveness of planning interventions [6–10].

Energy consumption data has a unique role in the larger open data ecosystem. Energy records are managed by a wide range of actors and groups, such as distribution system operators (DSOs), municipal tax agencies, national grid authorities, and private utilities. This is different from transport flows or land-use classifications. Each of these groups collects data for reasons that do not always fit with what spatial analysis needs. For example, taxation, billing, grid balancing, and regulatory compliance all require different structures, granularities, and levels of spatial attribution on records. Consequently, the analyst aiming to map energy consumption at the neighbourhood or building level often faces a paradox: the data are available, yet their format makes them analytically inaccessible without considerable preprocessing.

This paper offers a dual perspective on this challenge, utilising two ongoing research initiatives in urban settings in Italy and Portugal. The Italian case study focuses on Naples, which is one of the largest and most crowded cities in Southern Europe. The Siatel platform, a municipal fiscal database that connects utility billing records to individual user addresses, was used to get energy consumption data here. The dataset is very detailed; it has records for each user over several years and includes most of the city's residential and commercial buildings. But the addresses in Siatel are based on tax administration rules rather than topographic maps. They are often incomplete, not in the right format, or have systematic typographic noise that makes it impossible for automated systems to work with them. The main methodological issue is how to reliably assign geographic coordinates to hundreds of thousands of noisy administrative address strings. This is a problem of spatial integration.

The case study of Portugal faces a problem that is almost the opposite. The national DSO and regulatory authorities in Portugal make energy consumption data available through national open data portals. These datasets are combined into administrative units. The methodological challenge here is disaggregation: how to turn macro-level energy stats into reasonable estimates for smaller areas that can be used for neighbourhood-level research. Spatial interpolation, dasymetric mapping, and regression-based downscaling techniques need to be carefully adjusted to fit the shape of the city.

The comparison of these two datasets is highly analytically productive, as they represent the diverse challenges of working with real-world urban energy data. By comparing a high-resolution but spatially noisy dataset and a spatially clean but low-resolution dataset, this paper aims to demonstrate how data structure influences the choice of tools and, therefore, the quality and usability of the outputs. The value of this paper is not in

an exclusive technique but in a comparison of how open data constraints affect urban energy analysis. The comparison of these two datasets is analytically productive because they represent two non-equivalent pathways through which urban electricity indicators are constructed from real-world data infrastructures. By comparing a high-resolution but spatially noisy dataset and a spatially cleaner but lower-resolution dataset, this paper shows how data structure influences the choice of tools, the definition of the target variable, and the validity of subsequent comparison. The value of the paper therefore lies not only in the treatment of open data constraints, but also in demonstrating that cross-city urban electricity comparison is defensible only when the target variables are explicitly harmonized across distinct data infrastructures [11–13].

The following section presents a theoretical position on geocoding, spatial disaggregation, and analytical equivalence in the context of urban informatics and open data. Section 3 describes the materials and methods used in both case studies, while Sect. 4 presents results and comparisons between them. Section 5 concludes by dis-cussing the harmonization requirements that still condition robust cross-city analysis.

2 Theoretical Framework

Open data in urban governance has evolved over the past two decades. Initially, this topic generated considerable enthusiasm and interest; however, over time, there has been a growing recognition within the scholarly and policy communities that government data are often less readily accessible than previously assumed. This realisation has highlighted the challenges involved in data transparency and availability, indicating the need for more systematic approaches to improve accessibility. In particular, as Janssen et al. [10] showed, open data remains confounding for those who want to make use of it: technical barriers such as formats not compatible with each other, absence of APIs, legal barriers such as license limits or privacy issues, and institutional barriers such as fragmented mandates, competing stakeholder interests. This is particularly relevant to the energy sector, where the interests of distribution networks, regulators, local governments, and open-data users intersect on several issues.

Within this landscape, urban energy data are considered particularly sensitive forms of open data. Unlike land-use or transport records that describe the built environment at a broad scale, energy consumption records contain details about households, the premises they occupy, and citizens' yearly energy use (often seen as a proxy for annual income). However, full disclosure of this information is unlikely due to data protection rules such as the European General Data Protection Regulation (GDPR). In light of this balancing act between analytical usefulness and the protection of privacy, a number of EU member states release individualised data from their smart meters with strict anonymisation, whereas others release only aggregated spatial scales.

Within this regulatory environment, the characteristics of Italian and Portuguese energy data governance reflect wider national institutional contexts. In Italy, without a centralised open data infrastructure for residential energy use, researchers must often use administrative datasets such as tax or cadastral records. The Siatel platform, a shared access system operated by the Italian Revenue Agency (Agenzia delle Entrate), is one example of an open data system that aggregates tax-related utility information per person

but only provides spatial information in the form of free-text administrative addresses. In Portugal, the transition to smart meters and government demands for data transparency have led to more standardised open datasets, but at the cost of spatial resolution, where most publicly available data are collected at the postcode or administrative unit level.

Different dataset designs demand different strategies for managing open energy data and applying it in urban planning. The Italian case involves a detailed geocoding process to convert data collected by various actors (primarily for fiscal reasons) into usable datasets. Conversely, the Portuguese case requires dasymetric and disaggregation techniques. Different dataset designs demand different strategies for managing open energy data and applying it in urban planning. The Italian case involves a geocoding process that converts user-level administrative records into spatially usable observations. Conversely, the Portuguese case requires disaggregation and spatial translation techniques that reconstruct tract-level indicators from aggregated low-voltage records. In both situations, the resulting electricity indicators are constructed analytical objects rather than simple direct observations, and cross-city comparison therefore depends on harmonizing the meaning of both numerators and denominators across the two pipelines [14].

2.1 Geocoding

Geocoding addresses is, in practice, far more complicated than it might appear. The task, which consists of matching a piece of text describing a location to a set of geographic coordinates, becomes progressively harder as the quality of the input data deteriorates. When addresses are well formatted and the underlying gazetteer consistently covers civic numbers, automated systems work well. The trouble is that administrative datasets collected for purposes other than spatial analysis (e.g. tax records, utility bills, planning registers, etc.) almost never meet these conditions. Strings riddled with abbreviations, inconsistent street-type prefixes, typos inherited from manual data entry, or informal references to local landmarks that simply do not appear in any national address database may often be found.

Scientific literature and IT tools developed on geocoding quality have long distinguished between two types of error that pull in opposite directions. The first is location accuracy: how far the geocoded point ends up from the actual location. The second is the match rate: what share of records the system manages to geocode at all. What needs to be taken into account is that improving one may lead to worsening the other. In fact, accepting looser, approximate matches to include more records in the geocoded dataset will inevitably result in some locations being assigned to the wrong street or placed at a street centroid rather than the actual building entrance. Conversely, insisting on exact civic-number matches improves accuracy but leaves a nontrivial portion of the dataset unmatched.

Which trade-off makes sense depends heavily on how the data will actually be used. A building-level energy audit cannot tolerate a geocoded point that is two streets away; accuracy has to come first. A citywide map showing consumption patterns in each neighbourhood can accept some positional inaccuracy as long as coverage remains high, because missing entire neighbourhoods from the analysis is probably the most undesirable result.

Open-source geocoding engines, especially Nominatim, have become increasingly popular as cost-effective alternatives to commercial services such as the Google Maps Platform or ArcGIS World Geocoding Service [15, 16]. Nominatim uses the open source OSM database (with global coverage and community maintenance) that is very heterogeneous in terms of civic number completeness across regions and cities. To sum up, when comparing Nominatim to commercial services, while Nominatim performs reasonably well in densely covered urban areas (with OSM coverage), the accuracy of civic numbers has been inconsistently reported or poorly infrequently.

Fuzzy string matching is an approach to bridge the gap between noisy address inputs and reference gazetteers. Approaches such as the Levenshtein distance, Jaro-Winkler similarity, and the Token Sort Ratio from RapidFuzz can calculate the similarity of two strings after they are compared with respect to character additions, word ordering, and abbreviations. The Token Sort Ratio is particularly useful for address matching because it normalises the comparison by first sorting the tokens of each string alphabetically, making the score invariant to word order, an interesting property given the variable nature of Italian administrative addresses for street prefixes: via, viale, corso, piazza, etc.). For the present comparison, geocoding quality is not only a local data-cleaning issue. It also determines the spatial support of the Naples numerator and therefore conditions whether that numerator can later be aligned with indicators produced under a differ-ent infrastructure. In this sense, geocoding uncertainty enters the comparative prob-lem at the level of target construction rather than merely at the level of cartographic representation [17, 18].

2.2 Spatial Harmonization

Spatial harmonization poses an inverse but related problem. When electricity data are released for spatial units other than the official, most widely used administrative units, the analyst must infer how much of that demand should be allocated to meaningful spatial entities for urban analysis. The resulting values are not observed directly at the target scale; they require the analyst to produce (if absent, as is usually the case) a translation model whose assumptions become part of the indicator itself. Area weighting, dasymetric allocation, and morphology-based weighting are typical strategies for this purpose, but each one embeds a different account of how demand is distributed across urban space [19, 20].

This distinction is analytically important because a shared label, such as (kWh/year/household), does not, by itself, guarantee equivalence across cities. A tract-level value derived from direct residential user records is not immediately com-mensurate with a tract-level value reconstructed from mixed low-voltage demand and building-based allocation weights. For this reason, valid cross-city comparison requires an explicit harmonization framework that specifies numerator scope, denominator meaning, spatial support, temporal reference, and uncertainty reporting. Rather than treating Naples and Porto as if they supplied the same variable at different resolutions, the two cases must be aligned as members of a harmonized family of target variables, including at least social normalization tracks such as kWh/year/person and kWh/year/household, and a morphological normalization track such as kWh/year/m2.

3 Materials and Methods

This study adopts a methodological-comparative design in which the two case studies are first treated according to their own data constraints and are then evaluated against a common harmonization problem. The comparison is organized around five dimensions: source-data structure, target-variable family, denominator meaning, spatial support and translation logic, and uncertainty reporting.

3.1 The Italian Case Study: Naples

The Italian case study uses data from the Siatel 2.0 platform (*Sistema di interscambio anagrafico e tributario elettronico*), which is run by the Italian Revenue Agency (*Agenzia delle Entrate*). The 2005 Financial Law (art. 1, commas 332–334, Law 311/2004) gave the Municipality of Naples access to the platform. This law requires all utility suppliers to send annual consumption records to the national tax registry. The dataset for the Municipality of Naples contains 518,177 records for residential electricity users (persone fisiche) for the 2021 fiscal year. These records are divided into two groups: type 1 users who live at the supply address and type 2 users who live somewhere else. Each record has an annual consumption value in kWh, a billing month count, and most importantly, a free-text address string that shows the supply location as the user entered it when they signed up for the contract.

Two additional and ancillary datasets were combined to help with the spatial analysis. The civic address database was downloaded from the National Geoportal (Geoportale Nazionale) of the Ministry of Environment and Energy Security, which works under the INSPIRE Directive (D.Lgs. 32/2010). It has 107,478 georeferenced address points for Naples, based on data from 2011. The official street register (Stradario Ufficiale) from the open data portal of the Municipality of Naples was used as the official source for standardizing street names. It has 3,826 street entries.

The 2021 ISTAT Population and Housing Census provided socioeconomic data, such as the number of families living in the area and the number of homes that were occupied. This data was broken down to the neighbourhood level across the 30 administrative subdivisions of the municipality. The main methodological challenge is that the free-text address strings in Siatel don't match the standardised toponymic entries in the official geographic databases. This is mainly due to the fact that Siatel's address strings are formatted according to tax administration rules and often have typographic noise, non-standard abbreviations, and extra cadastral descriptors. Managing 518,177 records made it impossible to do everything by hand, so automated tools had to be used. The methodological pipeline created to address this issue and described in this section has three main steps, as outlined in Fig. 1 below: automatic string normalisation, fuzzy matching, and human-guided correction, which is performed iteratively.

The first methodological step is noise cleaning, which means removing elements that are not necessary for spatial identification, such as postal codes, floor indicators, staircases and internal unit references (e.g. piano, scala, interno, palazzo), and other cadastral qualifiers that people often include in billing addresses. After cleaning, the street type prefix and toponym are separated from the civic number. Records marked SNC (*senza numero civico*, or without a street number) require special handling. The

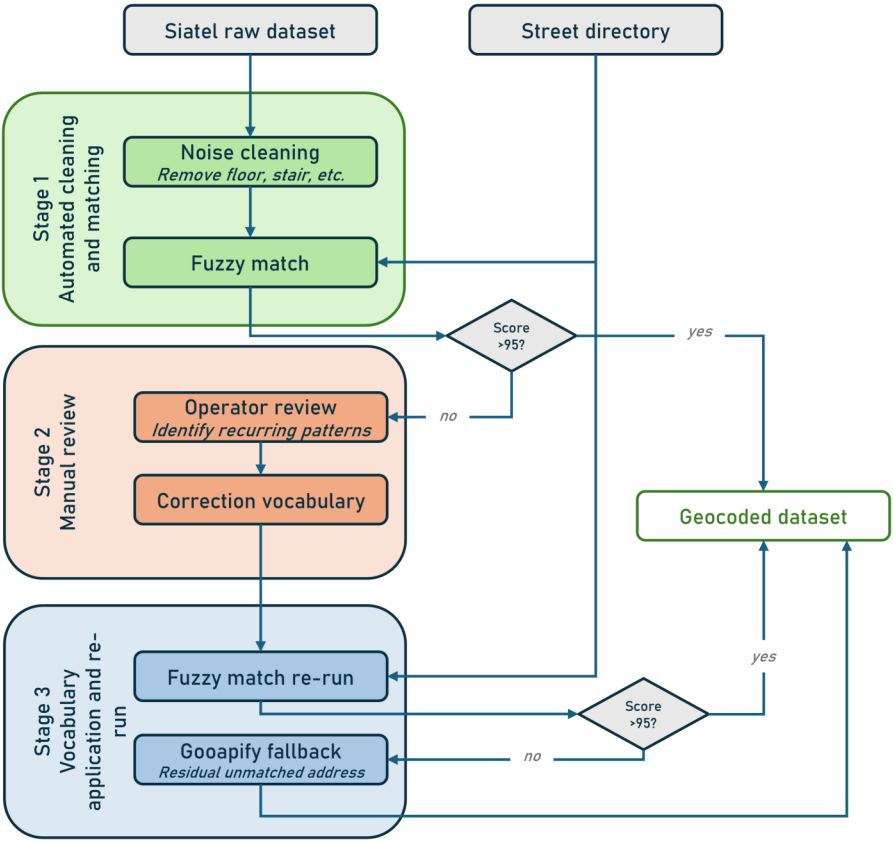


Fig. 1. Methodology workflow for Italian-based energy records.

cleaned address is then processed through a fuzzy matching phase that uses the Token Sort Ratio algorithm from the RapidFuzz library. This algorithm normalizes string similarity by sorting component tokens alphabetically before comparing them. This means that the matching score does not depend on the order of the words, which is useful because Italian street type prefixes (via, viale, corso, piazza, and their many shortened forms) can be used in different ways. Addresses that get a match score higher than 95 against the official street register are marked as "valid" and automatically given the right municipality and neighbourhood codes. Addresses that get a score lower than 95 are marked for "revision" and moved to the following methodological step.

The manual review consists of identifying recurrent patterns in the unresolved strings, such as local abbreviations of major thoroughfare names that are not included in the official register, the systematic inclusion of alternative street designations that reflect historical naming conventions, variations in civic number formatting (such as ranges like "15–17" or letter suffixes added to numbers), and strings that refer to buildings or complexes by name. A structured Correction Vocabulary, a mapping table that links problematic input strings to their confirmed equivalents, is used to record identified

corrections. The vocabulary is treated as a cumulative, version-controlled asset; new entries are added with each processing cycle, gradually reducing the amount of manual correction required in later runs.

The methodology has been designed to run iteratively, in order to maximise at each stage the number of strings recorded as "valid". Before the first step is run again, the Correction Vocabulary is applied as a pre-processing step to the complete original dataset. Additionally, a Substitutive Civic Map is created that, after validation against the official register at a 95% threshold, assigns the first valid civic number for a street to all SNC records connected to that street. A cache mechanism was put in place to control the computational cost of repeated fuzzy matching across a dataset this size: string comparisons that have already been calculated in a prior cycle are saved and retrieved directly, preventing redundant computations. After the corrected and re-normalised dataset has been processed through the entire first phase, the remaining "in-revision" records are sent to a final geoprocessing step that uses the Geoapify geocoding service to recover geographic coordinates for addresses that cannot be resolved against the municipal database, while "valid" records are written to the output.

The geocoded consumption data were combined with the civic address shapefile in QGIS after address correction was finished. There should be a clear explanation for the spatial aggregation unit selection. In this context, census tracts, the standard unit for Italian sociodemographic analysis, have two limitations: their geometry has changed over census rounds, making it impossible to compare historical series directly, and their boundaries often include street surfaces, which can result in address points being assigned to adjacent sections rather than the residential block they belong to. A 100-m hexagonal grid was adopted for the spatial aggregation of consumption data in Naples. This choice is justified by the need for spatial stability and time-invariance. Unlike census tracts, whose geometries are frequently modified between census rounds, the hexagonal grid provides a consistent unit for long-term longitudinal observation. Furthermore, this grid diameter minimises the risk of boundary effects where address points are erroneously assigned to adjacent streets, while maintaining a resolution sufficient to capture the block-level heterogeneity of the Neapolitan urban fabric. Spatial join was used to aggregate consumption values to the hexagonal cells by adding up all of the kWh values of address points that fell inside each cell. In order to compare energy behaviour across areas with various socioeconomic and morphological characteristics, a parallel aggregation to the 30 municipal neighbourhoods was also carried out.

The ISTAT census variables PF1 (resident families) and A2 (occupied dwellings) were used as denominators to calculate per-family consumption averages for the neighbourhood-level analysis. This produced a normalised indicator that takes into account variations in residential density across the city's extremely diverse urban fabric.

In the comparative framework developed in this paper, the Naples workflow is interpreted as the construction of an observed residential numerator that can subsequently be normalized using multiple denominator logic. The current Naples-side outputs therefore support at least a social normalization track based on resident families and occupied dwellings, while a person-based track remains contingent on the final alignment between available census denominators and the chosen spatial support.



Fig. 2. (a) Portuguese case study (Porto Municipality) – Census Tracts from 2021 census; (b) Buildings footprints and doors numbers (include streets names); (c) Postal codes polygons after combining doors numbers and street names as keys to merge postal codes raw text file; (d) Summary overview of geometric layers overlaps.

3.2 The Portuguese Case Study: Porto

The Portuguese case study addresses the inverse condition to Naples.

Porto, the second-largest city in Portugal, presents an urban context as complex as that of Naples but with distinct morphological drivers. With a total area of approximately 41.4 km^2 and a hilly topography, the city is characterised by a high-density medieval and 19th-century core surrounded by mid-20th-century expansions and contemporary suburban developments. Its Atlantic climate significantly influences residential heating and cooling demands. Unlike Naples, where user-level records were available, Porto's energy data is primarily governed by a transition toward standardised smart-metering at the postcode level, requiring a top-down spatial translation to identify vulnerable neighbourhoods.

The Porto workflow reconstructs annual low-voltage electricity indicators from observations organized at the CP7 postal-code level and reallocates them to the BGRI2021 census tracts. Because CP7 areas are not available as official analytical polygons, they are first inferred from matched door points and then spatially approximated through tessellation and clipping procedures. The resulting polygons serve as the origin support for allocating annual electricity totals to the target tract geography. Municipal building footprints, roof and ground elevation points, and terrain information are introduced to estimate built form and to derive allocation weights (Fig. 2).

Two allocation strategies are retained in parallel. In the first, electricity is redistributed from CP7 polygons to census tracts according to area-based intersection weights. In the second, the translation uses estimated building volume, derived from footprints and height information, so that the allocation reflects the internal distribution of built mass more directly. Retaining both strategies allows testing the sensitivity of the resulting tract-level indicators to alternative translation logics, rather than masking this uncertainty in a single output.

The Portuguese workflow produces a family of indicators rather than one tract-level metric. Let $((E_j^A)$ denote the annual electricity allocated to tract (j) under the area-weighted procedure and (E_j^V) The annual electricity allocated under the building-volume-weighted procedure. Let (N_j^{ind}) denote the number of individuals, (N_j^{hh}) the household or family denominator adopted for the tract, and (GFA_j) the estimated gross floor area. The resulting target family is given by

$$I_j^{A,ind} = \frac{E_j^A}{N_j^{ind}}, \quad I_j^{A,hh} = \frac{E_j^A}{N_j^{hh}}, \quad I_j^{A,GFA} = \frac{E_j^A}{GFA_j}$$

and

$$I_j^{V,ind} = \frac{E_j^V}{N_j^{ind}}, \quad I_j^{V,hh} = \frac{E_j^V}{N_j^{hh}}, \quad I_j^{V,GFA} = \frac{E_j^V}{GFA_j}.$$

In this configuration, the Porto-side numerator does not isolate residential demand with the same precision as the Naples dataset. It is better interpreted as mixed low-voltage electricity translated to the tract level under explicit spatial assumptions. For that reason, the Porto indicators are treated in this paper as analytically useful but only partially

commensurate with the Naples residential indicators until a target-level harmonization rule is specified.

3.3 Comparative Harmonization Logic

To establish a defensible Porto–Naples comparison, the two workflows are organized into a harmonized family of target variables rather than forced into a single nominal metric. The primary harmonization path is social normalization, which supports indica-tors such as (kWh/year/person) and (kWh/year/household). A complementary morphological path uses (kWh/year/m2) to retain a built-form-sensitive perspective for subsequent analysis. Across both tracks, comparability depends on making explicit five elements: the numerator scope, the denominator meaning, the spatial unit receiving the allocated value, the temporal reference attached to the observation, and the un-certainty introduced by geocoding or spatial translation.

Under this logic, a shared unit label is treated as a necessary but insufficient condition for comparison. A Naples indicator, normalized by resident families and derived from observed residential user records, is not analytically equivalent to a Porto indicator, normalized by family nuclei, if the numerator still includes mixed low-voltage demand or if the household definition differs institutionally. The comparative method, therefore, proceeds by identifying which indicators are already aligned, which are only partially aligned, and which require additional adjustment before any inferential cross-city modeling can be considered.

4 Results

4.1 The Italian Case Study: Naples

Naples is a densely populated Mediterranean city, with more than 900,000 residents concentrated within just 117 km^2, with over 7,700 inhabitants per km^2 in the most densely populated areas. Over 2,000 years of history, the city has developed through a long, layered process of urban growth, producing sharp contrasts between its historic core and its peripheral districts. The central areas, characterised by extremely compact medieval and early-modern fabrics, mid-rise buildings, narrow streets, and scarce open spaces, present complex spatial conditions shaped by centuries of densification and are subject to rigid heritage-preservation constraints. In contrast, the peripheral zones, although less dense, are often marked by fragmented urbanisation, larger open spaces, discontinuous building patterns, and, in some cases, major industrial or infrastructural enclaves. The city's overall structure is further conditioned by significant variations in altitude. These combined physical and functional characteristics, topographic fragmentation, limited developable land, heterogeneous settlement structures, and high demographic pressure make Naples a challenging context for urban planning, requiring attention to risk mitigation, mobility, heritage conservation, and equitable service provision. This complexity is also reflected in the city's energy demand: residential electricity consumption is strongly influenced by local urban form, socio-economic composition, and the morphological differences between dense central districts and more dispersed peripheral areas.

Recent studies emphasise the need for fine-scale spatial modelling to support mitigation and adaptation strategies, especially given the structural inefficiencies typical of compact Mediterranean cities and the heightened exposure to climate-related risks. The analysis presented in this study relied on data from the SIATEL 2.0 platform, which provides annual electricity and gas consumption data for all domestic and non-domestic users. Although highly detailed, the dataset contains addresses formatted according to administrative practices rather than cartographic standards, resulting in mismatches, outdated street names, and inconsistent numbering. To reconstruct the city's spatial energy demand, SIATEL data were harmonised and integrated with authoritative geographic information from the National Geoportal, including the municipality's civic-number database and ISTAT census-section boundaries, enabling each consumption record to be associated with a precise spatial entity through an extensive geocoding and preprocessing workflow.

The application of GIS methodology to the city of Naples produced the detailed geolocalization of residential electricity consumption, covering more than 518,000 domestic users. After the iterative correction and validation of addresses and the civic-numbers database, 93% of the SIATEL records were successfully geolocalised.

To represent the spatial distribution of energy data for this application, the census-section boundaries were not used, as they often include street segments and change geometrically over time. The final mapping was performed using a regular hexagonal grid (100 m). Using spatial join operations that accounted for the localisation of residential buildings, all validated consumption records were aggregated into grid cells, producing a detailed and uniform representation of the city's energy demand.

The resulting maps reveal clear spatial relation in the energy residential consumption: high-consumption areas are concentrated in the central and hillside neighbourhoods, particularly San Lorenzo, Pendino, San Giuseppe, Montecalvario, Avvocata, and Vomero. These areas feature dense residential fabrics, mixed land uses, and high occupancy rates, which collectively intensify electricity demand; low-consumption areas are mainly located in peripheral districts such as Chiaiano and San Pietro a Patierno, as well as in major industrial zones (Bagnoli, Zona Industriale), where large open spaces and lower residential densities lead to reduced domestic energy use (Fig. 3).

Additionally, the integration of residential consumption with demographic data from the 2021 ISTAT census enabled a comparative assessment across the city's 30 districts. The distribution of average electricity use per household aligns with expected behavioural and socio-economic patterns. High-consumption clusters in the historic centre (e.g., San Lorenzo, San Giuseppe) do not solely reflect high occupancy, but also the structural inefficiencies of older, uninsulated building stocks. Conversely, the lower consumption observed in peripheral industrial zones and recently developed areas like Chiaiano reflects both lower residential densities and, in some cases, the presence of more modern building materials that reduce the demand for active thermal regulation. Such distribution of average electricity use per household broadly aligns with expected behavioural and socio-economic patterns, reinforcing the reliability of the methodology and highlighting the analytical value of combining real consumption data with socio-demographic indicators.

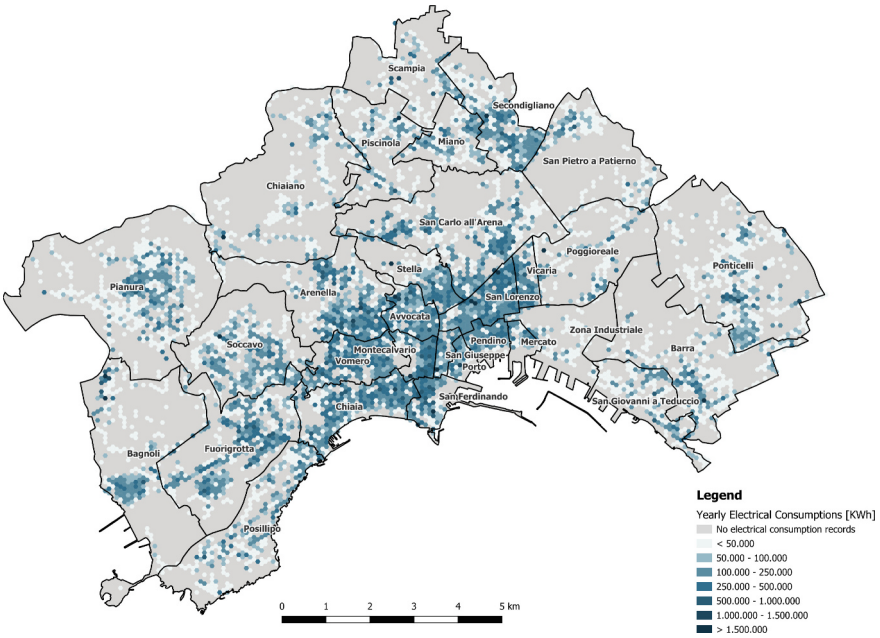


Fig. 3. Naples kWh/household distribution.

Overall, the Naples case study demonstrates that detailed user-level datasets, when properly harmonised, cleaned, and integrated within a GIS framework, can provide significant insights into the spatial heterogeneity of urban energy demand. The methodology confirms the importance of addressing data-quality constraints and adopting flexible, high-resolution spatial units to support effective and context-sensitive energy planning.

Overall, the Naples case study demonstrates that detailed user-level datasets, when properly harmonised, cleaned, and integrated within a GIS framework, can provide significant insights into the spatial heterogeneity of urban energy demand. The methodology confirms the importance of addressing data-quality constraints and adopting flexible, high-resolution spatial units to support effective and context-sensitive energy planning. At the same time, the results for the Italian case study clarify that achieving a high geocoding rate locally does not, by itself, solve the comparative problem, because the residential numerator produced here still needs explicit alignment with the numerator scope, denominator logic, and spatial support adopted in Porto.

4.2 The Portuguese Case Study: Porto

The Porto workflow yields tract-level electricity indicators via spatial translation rather than by directly geocoding user records. Its first result is therefore methodological: annual low-voltage electricity observed through the postal-code infrastructure can be reassigned to BGRI2021 census tracts using explicit, replicable weighting rules. This produces a sub-municipal analytical surface suitable for urban comparison while also preserving the distinction between the origin and target geographies.

A second result concerns indicator plurality. Because the Porto numerator is translated under both area-based and building-volume-based weights, the workflow produces parallel tract-level values instead of a single definitive allocation. This parallel structure makes the sensitivity of the final indicator to the assumed relationship between electricity demand and urban form visible. The Portuguese case thus yields not a single target variable but a family of plausible targets, bounded by different translation logics.

A third result concerns normalization. By combining translated electricity totals with census counts and estimated gross floor area, the Porto workflow produces socially normalized indicators such as kWh/year/person and kWh/year/household (Fig. 4), together with a morphological normalization track expressed as kWh/year/m2. These outputs make Porto analytically rich for intra-urban assessment, but they also show that the substantive meaning of the indicator depends on whether the analyst is studying demographic burden or built-form intensity.

The Porto-side limitations remain structurally important. The numerator reflects mixed low-voltage demand rather than a purely residential load, the origin polygons are inferred rather than official, and the built-form denominators depend on estimated gross floor area rather than directly observed architectural data. The resulting indica-tors are therefore suitable for comparison only when these constraints are explicitly retained in the interpretation of the target variable.

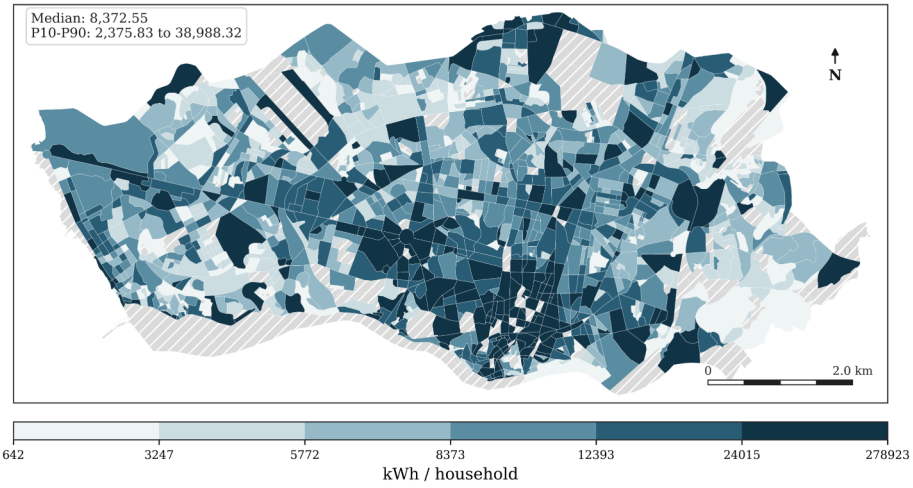


Fig. 4. Porto kWh/household distribution.

4.3 Comparative Synthesis

Taken together, the two case studies show that both cities can support sub-municipal electricity analysis, but they do so through different data-construction pathways. Naples begins with user-level residential observations and must solve spatial instability in address strings. Porto begins with aggregated low-voltage observations and must solve the translation from one geography to another. The apparent convergence of both

workflows at tract- or neighborhood-scale indicators should therefore not be mistaken for full analytical equivalence.

What is already comparable is the possibility of developing socially normalized indicators to address questions about demographic variation in urban electricity demand. Both cases also confirm strong intra-urban heterogeneity, indicating that city averages conceal relevant differences in morphology, occupancy, and social structure. What remains only partially comparable is the precise meaning of the numerator and, to a lesser extent, the meaning of household-based denominators across the two national statistical systems.

The main methodological result of the merged analysis is therefore not a single common metric, but a harmonized family of target variables. Within that family, some indicators can serve as primary comparison tracks, while others serve as complementary tracks. Socially normalized indicators are the most immediate candidates for cross-city work, provided that the numerator scope and denominator definitions are carefully documented. Morphologically normalized indicators remain valuable, but they should be interpreted as a complementary analytical track because they depend more directly on estimated built-form measures on the Porto side.

5 Discussion and Conclusions

The combined evidence from Naples and Porto indicates that open energy data be-comes useful for urban comparison only after a substantial phase of methodological transformation. In Naples, that transformation is dominated by geocoding, address normalization, and record recovery. In Porto, it is dominated by spatial translation, allocation weighting, and denominator construction. These are not peripheral technical adjustments. They define what the electricity indicator means and, therefore, set the conditions under which comparisons are valid.

For this reason, the paper's main conclusion is that cross-city comparisons and modelling of urban electricity are valid only when target variables are explicitly harmonized across distinct data infrastructures. A nominally shared unit such as (kWh/year/household) is insufficient if one city measures residential user-level demand and the other reconstructs mixed low-voltage demand through polygon translation. The comparison becomes defensible only when the analyst specifies the numerator scope, the denominator logic, the receiving spatial unit, the temporal frame, and the uncertainty structure attached to each indicator.

This conclusion does not diminish the value of either dataset. On the contrary, it clarifies how each one should be used. The Naples dataset is strong for recovering the spatial heterogeneity of residential electricity demand once address instability is resolved. The Porto workflow is strong for generating tract-scale indicator families from aggregated infrastructure data and for testing how different translation assumptions affect the resulting urban profile. Their joint value lies in showing that transferability research must begin from analytical equivalence, not from the availability of similarly named variables.

The methodological frameworks presented here offer direct support for municipal Sustainable Energy and Climate Action Plans. For the City of Naples, the 93% geocoding

success rate enables hotspot mapping at the street-segment level, allowing planners to prioritise building-envelope retrofitting subsidies for specific high-consumption blocks. In Porto, the top-down approach provides a sensitivity analysis tool for policy-makers; by comparing area-weighted vs. volume-weighted demand, technicians can more accurately estimate the impact of urban morphology on energy poverty. These data-driven insights transform administrative noise into actionable intelligence, ensuring that interventions are targeted where the intersection of social vulnerability and energy inefficiency is most acute.

Failures persist in both methodologies, which must be considered for sound decision-making. In Naples, the 7% of geocoding-resistant records were randomly spread, indicating no systemic geographic bias in the final spatial indicators. In Porto, sensitivity analysis showed that switching from area-based to building-volume-weighted allocation caused an average variance of 12% in tract-level estimates, emphasizing the need to use morphological proxies when user-level data is unavailable.

Future work should therefore focus on completing the harmonisation process rather than merely extending the modelling stage. On the Naples side, this includes finalising the denominator choice for the primary comparison track and documenting residual geocoding uncertainty at the selected support. On the Porto side, it includes clarifying the extent to which mixed low-voltage demand can be aligned with a residential comparison target and quantifying the sensitivity of the results to the chosen allocation rule. Only after these steps can Porto and Naples sustain a robust comparative modelling framework.

Acknowledgments. The authors would like to thank the Ufficio Tributi of Comune di Napoli for granting access to the Siatel database and for their support in the data retrieval and processing phases. The cooperation of the Naples municipal administration was essential to the conduct of this research.

References

1. Feitosa, F.O., Batista, P., Marques, J.L.: How to assess spatial injustice: distinguishing housing spatial inequalities through housing choice. Cities **140**, 104422 (2023). https://doi.org/10.1016/j.cities.2023.104422
2. Ascione, L., Gargiulo, C., Guida, C.: Adaptation and energy saving through urban green spaces in climate action plans: the experiences of 20 global cities. TeMA – J. Land Use, Mobility Environ. **2**, 31–46 (2025)
3. Gargiulo, C., Guida, C.: Climate-proofing cities: a focus on vulnerability indices. Int. Planning Stud. 1–19 (2026)
4. Weaver, T.: Urban crisis: the genealogy of a concept. Urban Stud. **54**(9), 2039–2055 (2017)
5. Carpentieri, G., Gargiulo, C., Guida, C., Zucaro, F.: Beyond city average: spatial analysis to support energy-saving interventions in urban planning. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., Rocha, A.M.A., Faginas Lago, M.N. (eds.) ICCSA 2024 Workshops. LNCS, vol. 14818, pp. 1–13. Springer, Cham (2024)
6. Long, Y., Liu, L.: Transformations of urban studies and planning in the big/open data era: a review. Int. J. Image Data Fusion **7**(4), 295–308 (2016)
7. Wang, C., Yin, L.: Defining urban big data in urban planning: literature review. J. Urban Planning Dev. **149**(1) (2022)

8. Barbosa, L., Pham, K., Silva, C., Vieira, M.R., Freire, J.: Structured open urban data: understanding the landscape. Big Data **2**(3), 144–154 (2014)
9. Östh, J., Nørgaard, N.F., van der Laag-Yamu, C.: The role of delineation for decision making and urban governance. In: Džunić, M., Östh, J., Muratori, S. (eds.) Advancing Urban and Local Governance in Western and Transition Europe. Local and Urban Governance. Springer, Cham (2026)
10. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Inf. Syst. Manag. **29**(4), 258–268 (2012)
11. Marie, P.: Using geocoded data in maps. In: Modern Web Cartography. Apress, Berkeley, CA (2026)
12. Pérez, V., Aybar, C.: Challenges in geocoding: an analysis of R packages and web scraping approaches. ISPRS Int. J. Geo Inf. **13**(6), 170 (2024)
13. Fichera, A., Frasca, M., Palermo, V., Volpe, R.: An optimization tool for the assessment of urban energy scenarios. Energy **156**, 418–429 (2018). https://doi.org/10.1016/j.energy.2018.05.114
14. Tavakkoli, S., Macknick, J., Heath, G.A., Jordaan, S.M.: Spatiotemporal energy infrastructure datasets for the United States: a review. Renew. Sustain. Energy Rev. **152**, 111616 (2021). https://doi.org/10.1016/j.rser.2021.111616
15. Hong, T., Chen, Y., Luo, X., Luo, N., Lee, S.H.: Ten questions on urban building energy modeling. Build. Environ. **168**, 106508 (2020). https://doi.org/10.1016/j.buildenv.2019.106508
16. HosseiniHaghighi, S., de Uribarri, P., Padsala, R., Eicker, U.: Characterizing and structuring urban GIS data for housing stock energy modelling and retrofitting. Energy Build. **256**, 111706 (2022). https://doi.org/10.1016/j.enbuild.2021.111706
17. Zandbergen, P.A.: Geocoding quality and implications for spatial analysis. Geogr. Compass **3**(2), 647–680 (2009). https://doi.org/10.1111/j.1749-8198.2008.00205.x
18. Helderop, E., Nelson, J.R., Grubesic, T.H.: 'Unmasking'masked address data: a medoid geocoding solution. MethodsX **10**, 102090 (2023). https://doi.org/10.1016/j.mex.2023.102090
19. Geng, G., et al.: Impact of spatial proxies on the representation of bottom-up emission inventories: a satellite-based analysis. Atmos. Chem. Phys. **17**, 4131–4145 (2016). https://doi.org/10.5194/acp-17-4131-2017
20. Zheng, B., et al.: Resolution dependence of uncertainties in gridded emission inventories: a case study in Hebei, China. Atmos. Chem. Phys. **17**, 921–933 (2016). https://doi.org/10.5194/acp-17-921-2017