

Article

Clustering Open Data for Predictive Modeling of Residential Energy Consumption across Variable Scales: A Case Study in Andalusia, Spain

Javier García-López , Samuel Domínguez-Amarillo  and Juan José Sendra 

Instituto Universitario de Arquitectura y Ciencias de La Construcción, Escuela Técnica Superior de Arquitectura, Universidad de Sevilla, Av. de la Reina Mercedes, 2, 41012 Sevilla, Spain; javigalo@us.es (J.G.-L.); sdomin@us.es (S.D.-A.)

* Correspondence: jsendra@us.es

Abstract: The energy budget of households, linked to residential energy consumption (REC), serves as a critical indicator of quality of life and economy trends. Despite the lack of widely available accurate statistics at regional or smaller scales, they are of crucial interest for a better understanding of the features influencing REC and its impact on energy poverty, wellbeing, and the climate crisis. This research aims to present a new information model for predictive parameters and REC forecasting through an innovative use of available open data. Geoprocessing, data mining, and machine learning clustering algorithms were applied to open datasets of location, population, and residential building stock parameters highly correlated with their REC, on the ensemble of 785 municipalities of Andalusia, Spain. The model identified 65 clusters of towns sharing the same potential REC, with 73% of the population concentrated in 10 of these. The resulting data-driven bottom-up model of provincial REC had a mean absolute error of only 0.63%. Furthermore, it provided the territorial distribution, with local resolution, of the identified clusters of cities with similar characteristics. This methodology, with a flexible regional- to city-scale analysis, provides knowledge generation that offers numerous practical applications for energy policy planning. Its future implementation would assist stakeholders and policymakers in enhancing the performance and decarbonization of the residential building stock.



Citation: García-López, J.; Domínguez-Amarillo, S.; Sendra, J.J. Clustering Open Data for Predictive Modeling of Residential Energy Consumption across Variable Scales: A Case Study in Andalusia, Spain. *Buildings* **2024**, *14*, 2335. <https://doi.org/10.3390/buildings14082335>

Academic Editors: Yu Huang, Siwei Lou and Yukai Zou

Received: 27 June 2024

Revised: 21 July 2024

Accepted: 25 July 2024

Published: 28 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: Andalusia; bottom-up model; clustering; energy model; open data; residential energy consumption

1. Introduction

In the European Union, the energy for residential building operation is responsible for a significant share of carbon emissions, as well as for the overall quality of life of citizens. Buildings in Spain account for 30% of these carbon emissions, according to the most recent consolidated report [1], with the energy consumption of residential buildings representing 18% of final national energy use. However, compared to transport or industry, this is still considered a ‘fuzzy sector’. This sectorial energy analysis, based on building stock features and national or regional surveys, statistics, and studies [2], often fails to include the characteristics of cities and towns. Nevertheless, there is growing interest in the distribution of energy use across the territory at different scales, despite the scant information available on energy use in residential buildings at district, town, or sub-regional scales [3].

Residential energy consumption (REC) is described as the result of the relationship between energy demand and the performance of the technical systems found in dwellings and how they are used, depending on household features [3]. However, energy consumption is also linked to multiple social, economic, cultural, environmental, physiological, and psychological aspects involving the users [4,5]. User profiles, therefore, determine the associated demand, the characteristics of the systems and equipment available, and in turn, the resulting energy consumption [6].

Residential energy consumption is broadly included in surveys and statistics worldwide, such as the Residential Energy Consumption Survey (RECS) in the USA, the Energy Follow-Up Survey (EFUS) in the UK, and all those referenced in the Manual for Statistics on Energy Consumption in Households (MESH) in the EU [3]. Studies about energy consumption in existing households, such as that carried out in China [7] at national and regional scales, describe the divergent evolution during the last two decades of urban and rural areas and its influencing drivers, including population and per capita consumption. However, as seen in the SPAHOUSEC study on the energy use of the Spanish residential building stock [8–10], surveys and statistics are often not available at sub-regional or town scales. The lack of adequate scales and detailed information on energy end use [11,12], with some exceptions [13], makes it difficult to study its nature and distribution across the territory, crucial for assessing a distributed energy production and consumption model.

In addition to the adequate scale, the quality and reliability of REC statistics are also major concerns, given that instances of misestimation in energy end use application statistics, as described for heating and cooling energy use in the case of southern regions in Spain [14,15], were also identified.

While open access to energy-related data is usually limited, available data on energy efficiency in buildings are also scarce, as described in the context of the indicator framework for the European EPBD Directive proposal in Spain and Italy [16]. In this study, only 35% to 40% of data were available, whereas the share of data from open sources ranged from 36% (Spain) to 90% (Italy) of total indexes. This demand for information on the energy efficiency of buildings continues to grow and must be met if global decarbonization targets are to be achieved.

In the context of the information society, quality open data are a valuable resource. This key tool for generating knowledge and returning value to society must be accessible and interoperable. Thus, the use of open data is a fundamental resource for pursuing sustainable development goals and informed decision-making in energy and urban planning [17]. This goal is featured in EU-funded projects, such as HOTMAPS, PLANHEAT, and Heat Roadmap Europe (HRE4). Web energy-mapping tools [12,18,19] currently being developed demonstrate the potential of this tool for local authorities, energy and urban planners, and ESCOs. Here, open-source datasets are shown at different scales: country level, regions, provinces, and municipalities, down to the hectare level (250×250 m grid).

Based on this, and on the literature, it can be concluded that the use of open data on building energy is gaining traction and is currently driving knowledge generation. However, nowadays, non-open data sources are a common resource owned and exploited by utilities in several forms [20].

According to reviews of the literature, multiple methods, such as the top-down or bottom-up approaches for the estimation of final energy use, can be used to forecast REC at different scales [21]. These technique names refer to the hierarchical position of data inputs in relation to the housing sector as a whole. The HOTMAPS study follows a top-down approach, showing heat density distribution at the hectare level from province population and gross floor area parameters, resulting in maps of space heating demand per gross floor area. In 2012, several authors [22] classified prediction models for building energy consumption following engineering, statistical, and even artificial intelligence (AI) methods. It was thus concluded that rapidly developing AI methods could provide alternatives or breakthroughs in the prediction of building energy consumption. In this respect, data-driven approaches, as described in recent literature [23–28], incorporate prediction and classification functions for building energy analysis, combining multiple AI and machine learning (ML) techniques. Data-driven models can, therefore, be oriented either to estimating the energy consumption of buildings (prediction models) or to profiling the consumption pattern (classification models).

The application of different bottom-up approaches for energy modeling found in the literature includes a wide range of examples from the last decade. Novel bottom-up physics-based models, such as urban building energy modeling (UBEM) [29], have been applied to

the cases of Rotterdam [30,31], Boston [32], Seville [33], Jaén [34], and Dublin [35,36]. These models, which consider the construction and engineering features of multiple interacting buildings and their thermal simulation, were used to estimate energy use in buildings from district to urban scales.

As an alternative to UBEM models, bottom-up statistical energy models based on GIS tools and statistics are described for Lisbon [37], New York [38], Turin [39], Seville [40], and Madrid [41,42] at both urban and district scales.

Recently, urban energy modeling (UEM), which includes urban services, was tested in Ahmedabad [43], providing an estimation of energy use from district to urban scales. In addition, an energy model for Singapore [23] provides a metropolitan-scale analysis of the effect of urban, environmental, and behavioral features on commercial building stock, predicting consumption accordingly. Smart-E is a UEM model for electricity use in dwellings from district to country scales, with time steps of only 10 min, validated with the annual electricity consumption of ten French cities [11]. Another recent study developed in France modeled real consumption for the entire residential building stock, calibrated with feature datasets for national buildings, and included occupant characteristics in order to reflect their behavior [44]. A similar engineering bottom-up model for estimating heating and cooling consumption of the whole residential building stock in the city of Chongqing used archetypes and BEM modeling validated with metered data [45].

However, there are two primary research problems with predictive models with respect to the scale and availability of REC datasets. Bottom-up regional energy models are scarce, as most models do not exceed the urban limits at territorial or regional scales. Furthermore, based on previous research, few energy models provide descriptions of the local environment or occupant features. Flexibility in the scale of application is also reduced, and frequently limited to a single scale: building, district, city, or national [46]. Actual REC data statistics under regional scale (cities, towns, or districts) are scarce and with limited access.

This research aims to reveal the ‘topography’ of residential energy use distribution within a given territory. To this end, as well as offering a novel information model for the predictive parameters (based on open data), an innovative open-access methodology is presented. These two hypotheses are considered to counter the limited availability of data at a suitable scale and quality for regional and urban energy planning. For the sake of actual application, a case study for Andalusia, in the south of Spain, is developed at the regional scale.

In view of the main research challenges outlined, this work introduces an innovative bottom-up energy model. This model relies on open data and processes for implementation, replication, adaptation, and periodic updates across varying scales in order to estimate REC. It also takes into account the features of locations, occupants, and dwellings for the municipalities analyzed (cities, towns, villages, etc.). Therefore, the estimation of potential REC of this model considers the combination of the features for the categories of location (L), occupancy (O), and dwelling (D).

The scale of application ranges from single municipalities to groupings into counties, provinces, regions, countries, etc., as well as clusters of municipalities that share common profiles, even if they are not physically close together. This flexible scale, named as the ‘regional–municipal scale’, considers the analysis of the potential REC of population centers (cities, towns, and smaller municipalities) within a single region.

Assuming, as is often the case, that a sole municipality is the minimal governance structure with environmental, social, housing, and energy competences, the regional–municipal energy model on offer can be of value to the energy-planning policies covering all these scales.

By implementing this model, the potential REC in a given territory can be mapped through the use of predictive models utilizing open datasets. The new information and knowledge obtained thus increase the public value of this model, especially in terms of

energy planning and decision-making, as defined in the data–information–knowledge hierarchy (DIK) described in the literature [47].

In summary, this study relies on a data-science clustering process for a set of towns and cities, classified by potential REC, considering their predictive REC parameters, to build a bottom-up energy model at the territorial scale.

Following this introduction, Section 2, ‘Methodology—Theory and Calculation’, includes the theoretical background of the proposed information model and the case study workflow explanation. An initial dataset correlation analysis for feature selection is followed by the dataset sources and description, geo-processing, preparation, clustering, and modeling for allocating potential REC to each municipality.

Section 3, ‘Results and Discussion’, introduces the selection of parameters as a result of a correlation analysis. The results of the clustering process are then analyzed and contrasted with current data for the individual information categories: location, occupants, and dwellings, in order to establish a ranking of potential REC among the clusters found in Andalusia. A final ‘cross-clustering’ process is then applied to find municipalities sharing the same L, O, and D clusters and, thus, the same potential REC. This is completed with a sample application in order to identify cities deserving priority funding for energy interventions. Finally, methodology and results are validated through the construction of a predictive model for REC at the provincial scale, compared with real consumption data.

The final section of the paper, Section 4, ‘Conclusions’, provides a summary of the principal findings of this research in countering the scarcity of data relating to energy consumption in residential buildings with bottom-up modeling based on open data. The main hypothesis has been confirmed using the case study in Andalusia, along with a data-science process from data to knowledge generation. Implementation of the methodology and future research developments are also described at the end of the manuscript. Lastly, Appendix A, ‘Dataset links’, includes the results of the case study for Andalusia, with links to the L, O, and D datasets and cluster allocation of the municipalities, in GIS and XLSX format.

2. Methodology—Theory and Calculation

The proposed methodology, based on data science processes of data mining (DM) and machine learning (ML), involves feature selection, data acquisition, debugging, post-processing (normalization) description, and ML clustering classification techniques. It has been drawn up using public datasets, post-processed with GIS tools, DM processes, and ML clustering techniques. The workflow applied examined a case study for the dataset of municipalities within Andalusia to obtain a town-to-regional REC prediction model. This methodology combines theory and calculation and can be extrapolated easily to other cases in other territories.

2.1. Theory—LOD Information Model Proposal

For an initial approach to this analysis, a tailored information model is proposed for the study and characterization of the parameters involved in REC modeling at the regional–municipal scale. In this case, the REC predictive parameters were organized into three main data categories, considered as independent dimensions: location (L), occupants (O), and dwellings (D).

According to this model, represented in Figure 1, for any point within a region at (x,y) coordinates (e.g., UTM coordinates of a building, district, city, etc.), the combination of the L, O, and D parameter vectors determines the REC value for that position (Equations (1) and (2)).

$$REC(x, y) = \left| \overrightarrow{LOD}(x, y) \right| \quad (1)$$

$$\overrightarrow{LOD} = a\overrightarrow{L} + b\overrightarrow{O} + c\overrightarrow{D} \quad (2)$$

where:

- $\vec{L}(l_1, l_2, l_3 \dots l_i) =$ location parameter vector (e.g., elevation, solar radiation, etc.)
- $\vec{O}(o_1, o_2, o_3 \dots o_i) =$ occupant parameter vector (e.g., average household size, age, etc.)
- $\vec{D}(d_1, d_2, d_3 \dots d_i) =$ dwelling parameter vector (e.g., buildings' height, area, etc.)
- $a, b, c = L, O,$ and D weighting factors.

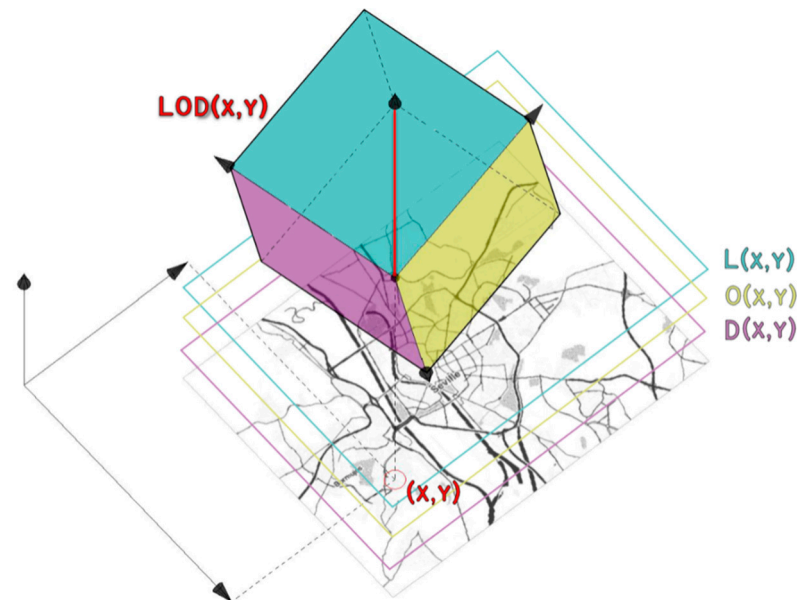


Figure 1. LOD data model for REC predictive parameters according to spatial distribution.

In Equation (2), the location (L) vector includes parameters such as elevation, solar radiation, etc. The occupants (O) vector includes coordinates such as household size, incomes, age, etc. The dwellings (D) vector describes height, age, number of dwellings, etc.

In this study, the weighting factors (a , b , and c), described in Equation (2), for the 3D composition of the LOD (x, y) vector in Figure 1, were considered equal to one, as in an isometric perspective. This initial configuration of L, O, and D vector composition can be tailored to any questions to be asked, and adapted through further calibration processes with actual energy use data following updated datasets available at the town scale. The model proposed presents a flexible framework ready to be adapted and calibrated with future growth of actual REC datasets. The weighting parameters can, therefore, be adjusted to assess the influence of the different L, O, and D predictive parameters, as well as to identify their sensitivity.

The bottom-up statistic REC model involves the analysis of a broad range of statistics, monitoring, and survey datasets. Given the recent development and potential for forecasting building energy performance [25], DM and ML techniques, including artificial neural networks, support vector machine, Gaussian process, or clustering algorithms, should be applied extensively. To this end, a workflow based on the OSEM—Obtain, Scrub, Explore, Model, iNterpret—process is applied [48,49]. This process, selected as the base for the workflow and widely adopted in data science, serves as a guideline to extract meaningful information and carry out data-driven decisions. The OSEM procedure is, in fact, an evolution of the KDD (knowledge discovery in databases) workflow. KDD consists of a nine-step sequence with possible backtracks, which starts with learning the application domain and ends with using discovered knowledge in practice [50]. The proposed adapted OSEM workflow aims to generate new information and identify common features amongst the municipalities analyzed. Thus, the workflow diagram shown in Figure 2 describes the stages, tasks, and results, from raw data to the final regional–municipal REC model validation stage. Columns A, B, and C in Figure 2 represent KDD/OSEM stages, tasks, and results, while rows 1 to 12 represent the direction of the process.

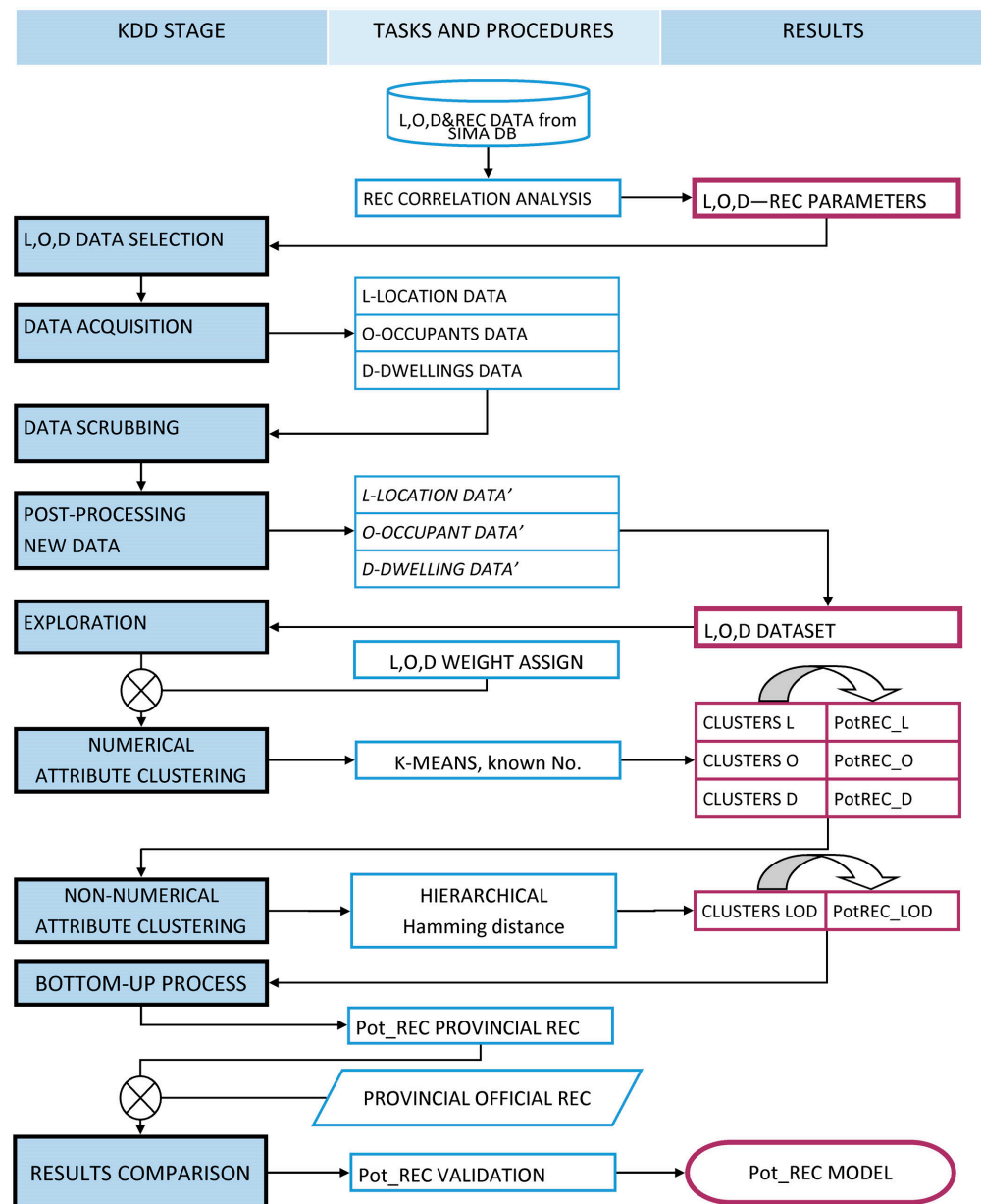


Figure 2. OSEMN municipal and regional REC model workflow.

This study focused on the residential energy use of households located in primary homes, since consumption in second homes in Spain is not considered significant [3,8,51]. The requisites for data source and predictive parameter selections for the regional to town REC information model data should be free, open, and publicly available from trusted sources. Such data must also have been updated prior to 2020 to avoid the likely pandemic bias effect, given that lockdowns and changes in social dynamics in REC evolution deserve their own dedicated study [52,53]. The datasets chosen, which must be available at urban or suburban scales, should also have a medium to high correlation with REC. Finally, the size of the predictive dataset adopted should ensure balance between model accuracy and complexity.

As REC is widely but discontinuously spread over a territory, it can be measured at different scales, including buildings, quarters, districts, cities, regions, etc. Parameters included in the REC data model from town to regional scales should not be affected by the size of the city or town in question. Therefore, parameters used should be normalized in terms of units

per inhabitant, households, area, percentage, etc. In many cases, this extra requirement means that raw data must be processed to obtain the new normalized parameters.

2.2. Calculation—Case Study

The region of Andalusia, considered highly representative of the Euro-Mediterranean context, was selected as a case study to put the approach described above into practice. This region in southern Spain has an extension of 87,599 km² and is similar in size to Serbia and larger than Croatia, two countries fully located within the same Euro-Mediterranean region. Although it has a predominantly Mediterranean dry climate (Csa, Csb), semiarid (BSh, BSk), arid desert (BWh, BWk), and even mountain (Dsa, Dsb, Dsc) Köppen–Geiger climates can also be found [54]. Its population of around 8.5 million inhabitants and approximately 3 million primary homes is spread over 13 cities with a population of over 100,000—including Seville (700,000 pop.) and Malaga (600,000 pop.)—and hundreds of smaller towns and villages. The case study, therefore, covers a wide range of locations and environmental, socioeconomic, and urban features for the building stock. To study REC, the location, population, and housing stock of the whole set of 785 municipalities which make up the Andalusian territory have been characterized and analyzed. As only dwellings used as primary or permanent residences in the individual towns are considered, the number of households registered had to be consulted. Information on building characteristics and dates of construction was obtained from the Spanish cadaster, using the INSPIRE statistical #250 m mesh published by IECA [55], while the population data recorded in the 2019 census provided information on socioeconomic features.

Energy predictive regional to building scale data models, such as SEMANCO [56,57], include parameters from different categories and origins. Reference literature establishes the selection of predictive parameters or features as a pre-requisite for the implementation of any ML model requiring extensive research [25].

The use of dataset correlation analysis based on the Spearman method is common in energy analysis [23,58,59], cadaster information [60], cartography generation [61], and the use of nighttime satellite images in demographic and socioeconomic studies [62]. Therefore, a prior correlation analysis of the Andalusian town and city datasets was carried out to study the overall relation of average electrical residential consumption (considering no other energy sources) and several other socioeconomic and demographic variables. Unlike other methodologies, such as the Logarithmic Mean Divisia Index (LMDI), applied to study the influence of features on a composite variable, such as energy consumption [63], correlation analysis offers a simple way to assess relationships between variables and trends' identification. Yearly Multi-Territorial Information System of Andalusia (SIMA) datasets [64] include a set of 123 parameters at the town scale. In this initial study, we considered a SIMA dataset, post-processed normalized parameters, and nighttime lighting (NTL) remote sensing, as shown in Figure 3.

The high number of parameters strongly correlated with residential electrical energy use justified its inclusion in the predictive LOD information model finally adopted. Correlation with REC, availability, and feasibility were considered in the selection of potential REC predictive parameters for the dataset of Andalusian municipalities.

In the following stage of this calculation, a DM process was developed to identify distribution patterns of the predictive REC parameters, which may have been previously unseen in the municipalities of Andalusia. The process began with data acquisition from available open sources, including the National Statistics Institute (INE), Andalusian Cartography and Statistics Institute (IECA), and the Spanish Cadaster. Raw data were post-processed using GIS tools (QGIS version 3.16.7) and geo-processes were applied to obtain the final L, O, and D dataset with the information vectors of the individual categories for each town.

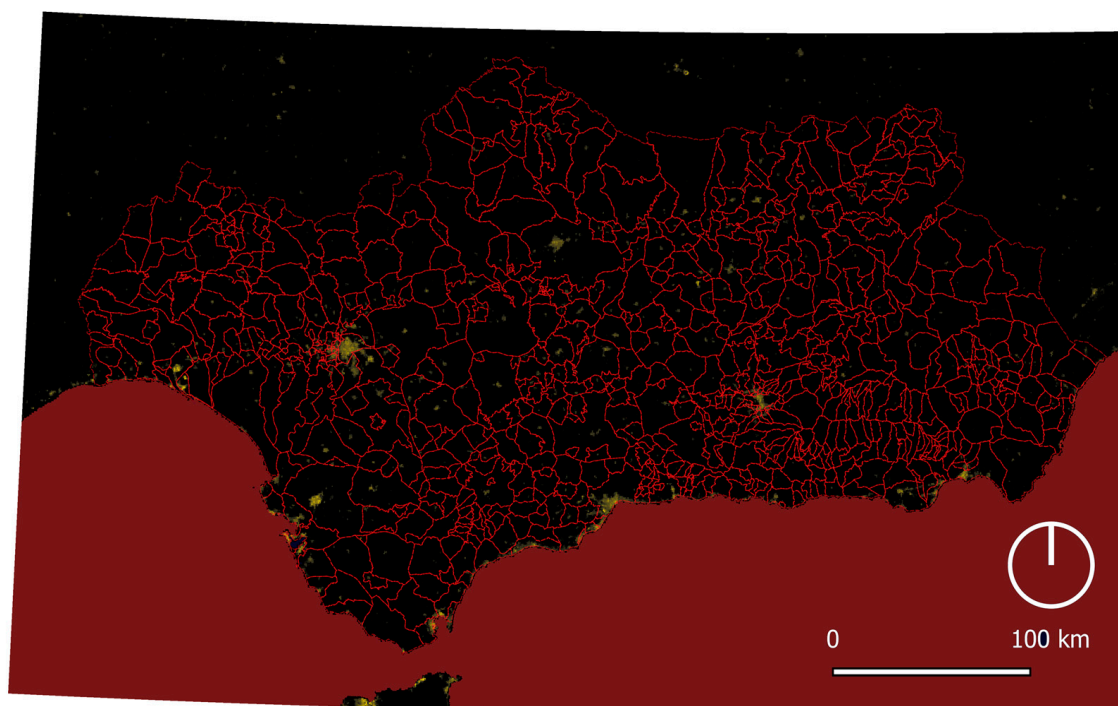


Figure 3. Nighttime lighting (NTL) over Andalusia. Image by the authors based on image background using Suoni NPP VIIRS and data from Miguel Román, NASA GSFC earthobservatory.nasa.gov.

According to the LOD data model described in Section 2.1, and Equation (1), zones with similar predictive REC parameters are expected to demonstrate similar REC patterns. The same LOD profile areas or populations might not be clearly identified or regularly distributed across the analysis region. Identifying these groups within the region datasets can be strongly improved with ‘clustering’, a ML technique that groups individuals with a greater number of similarities for the purposes of segmentation and discovery of patterns. There are numerous clustering techniques and algorithms, and in many cases, they are equally valid for the same data analysis process [65]. Clustering algorithms have also been applied to energy studies in the industry sector [66], local climate zone definition [67], user profile characterization [6,68], and building archetype identification [28,69,70]. Specific K-means or hierarchical methods were selected based on the study of related cases, such as the urban-scale building clustering process for heritage damage assessment [71] and energy predictive models in Geneva [70] as well as a district in Genk [24].

The ‘attribute-based clustering’ plugin for QGIS, developed by E. Kazakov, can be applied to vector layers with numeric or non-numeric attributes using a variety of Python ML clustering algorithms. Numeric values were grouped within a known number of clusters applying the ‘K-means’ clustering algorithm with weighting factors. When setting this K-means algorithm, weights (a, b, and c) had to be assigned to the numeric parameters within the individual categories studied (L, O, and D). As stated in Section 2.1, for this research, the weighting factors applied were set to 1 to ensure a neutral model, although later relevant aspects of the assessment could be fine-tuned or highlighted through further calibration. The forthcoming release of open data will improve knowledge generation through additional DM and ML processes.

The elbow method was applied to establish the optimal number of clusters to be sought in each category, considering the dataset features [68,72]. The optimal number of clusters desired for each class was set at five.

The following task included statistical analysis of the L, O, and D clusters identified, which helped in the assignment of a potential REC (PotREC) ranking using statistical inference. A final clustering action applied the Hamming-distance hierarchical algorithm, which is

particularly useful for categorical (non-numeric) values (e.g., classes or text fields) [73]. Thus, towns and cities from the same L, O, and D clusters and sharing the same potential REC were also identified. This final clustering or cross-clustering action helped to obtain the LOD cluster for each town in the dataset and, consequently, assign potential-LOD_REC (PotLOD_REC), as described for the city of Seville in Figure 4.

LOD Clustering

PotREC_LOD assignment. LOD cross category

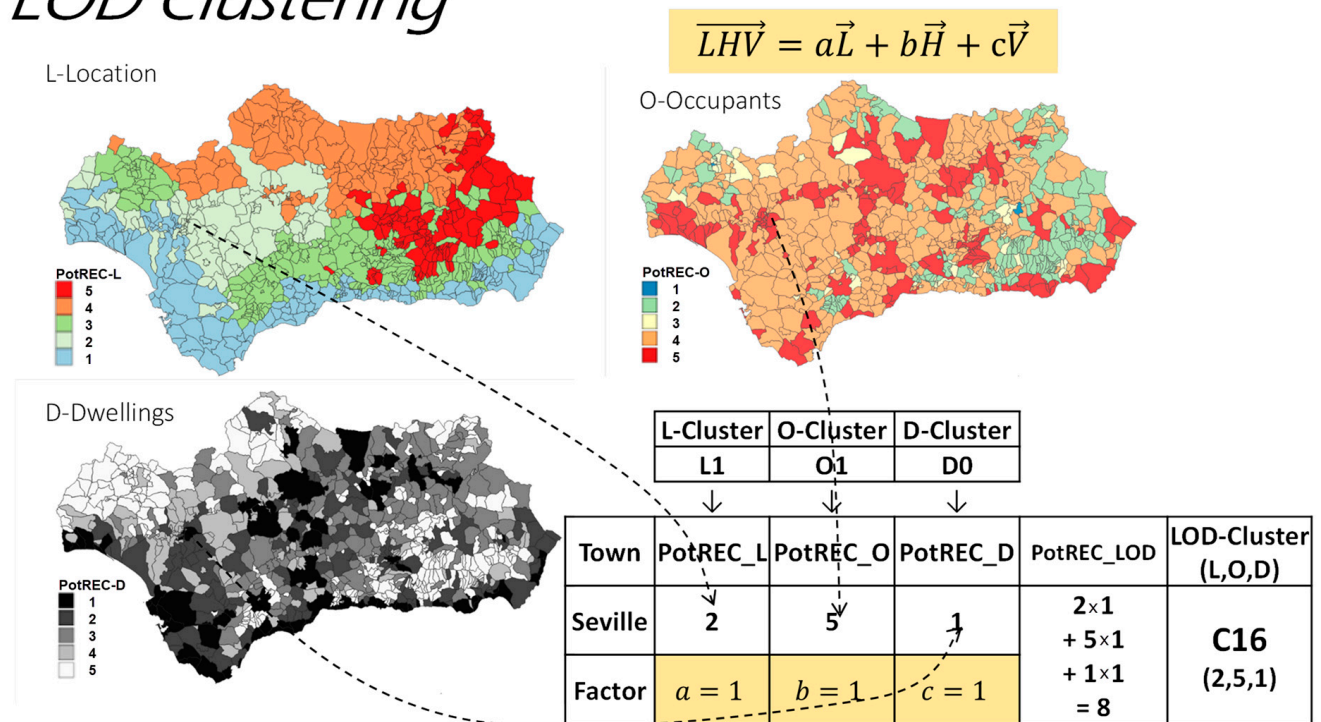


Figure 4. Allocation process for individual town LOD cross-clusters and LOD potential REC: Seville.

In the final stage of the OSEMN process, the clusters identified for Andalusia were assessed to establish a model for estimating REC at the provincial level. This was based on the computation of the stock of primary homes in the individual towns weighted using the PotREC_LOD index. The LOD model proposed was compared to two alternatives based on the distribution of regional consumption by population or by number of primary homes in each province.

The model described here relies on public datasets regarding location, inhabitants, and housing, updated until 2020, considering only primary homes and registered population within urban areas of municipalities of the region. Features selected for the model rely on the preliminary correlation analysis of available open datasets with actual electrical residential energy use at the town scale. It should be noted that the clusters identified exclusively describe municipalities within Andalusia. The PotREC parameter is a qualitative index of a single municipality's capacity for residential energy consumption. It does not provide a quantitative description of actual or expected consumption.

3. Results and Discussion

This study aimed to analyze a data-science clustering process for a set of towns and cities considering their predictive REC parameters, and establishing a classification by potential REC, as seen in Figure 5. The results shown here justify the methodology applied in relation to the objectives of the study.

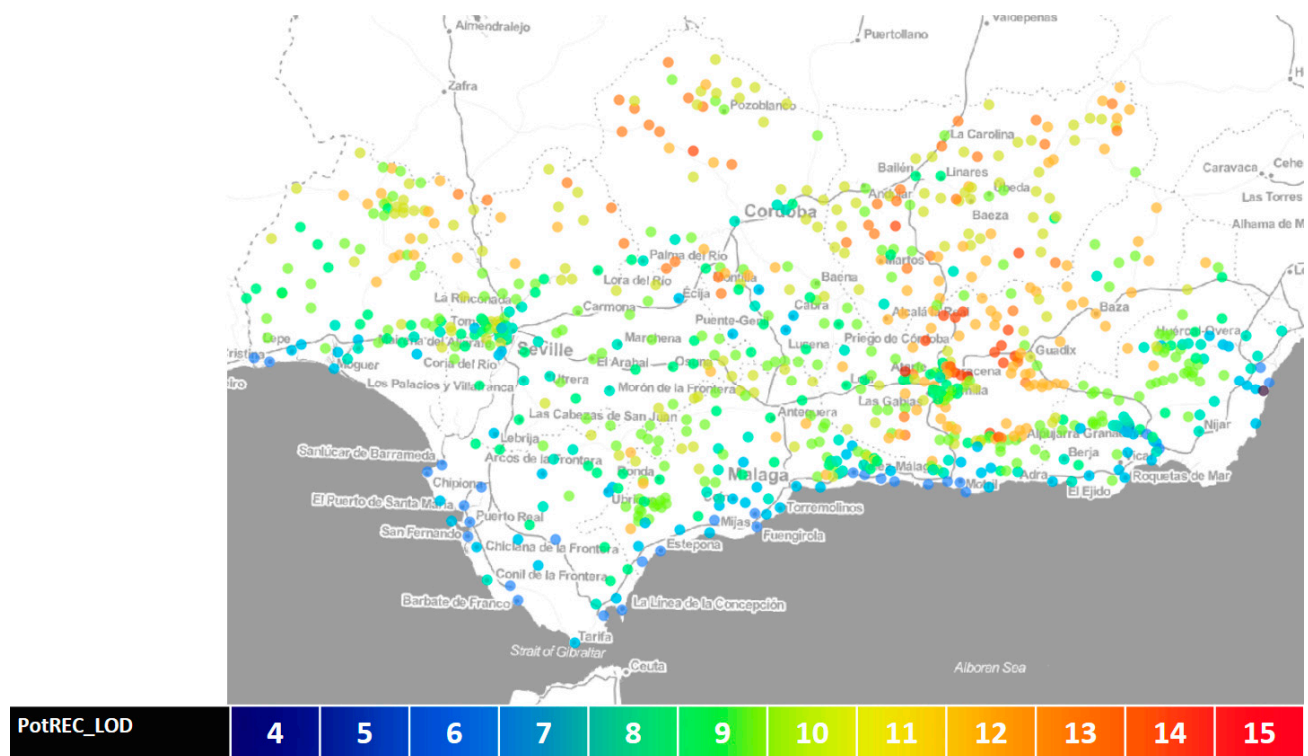


Figure 5. Distribution of potential REC across Andalusia.

The open datasets and methods combined in a novel methodology applied to this case study were successful in generating new and valuable information and knowledge relating to Andalusia.

3.1. Feature Selection

The first section of results presents an initial correlation analysis and the selection of REC predictive parameters for the LOD data model.

3.1.1. Correlation Analysis

An initial correlation analysis was conducted to select the set of predictive parameters for the LOD data model from available datasets at the town scale. This was used to describe the strength of the relationship with residential electricity consumption to build the model; however, it is not the model itself. For this purpose, we used the R ‘cor.test’ function, based on the Spearman method. The correlation coefficient and significance obtained for the 18 most representative parameters are summarized in Table 1. We used de Cohen’s guideline to describe the ρ correlation size effect [74].

The resulting positive or negative influence of the parameters on the REC was as expected and has already been documented in several studies [5,75].

Parameters closely correlated with electric consumption were those related to demographics, household size, income, age, and building size. It is also worth noting some other parameters’ correlation, such as municipality net population and its variation in the last decade, although these were not included in the LOD data model. Parameters linked to the urban environment, such as the heat island effect (UHI), nighttime lighting (NTL), the degree of urbanization, and average building height, were within a medium range of positive influence. A negative medium correlation was found for the proportion of single-family dwellings and the altitude of the municipality. Low correlated parameters include those related to climate, such as winter and summer climate severity. Remarkably, the town unemployment rate showed the lowest Spearman correlation coefficient, with no significance (p -value = 0.048).

Table 1. Spearman’s ρ correlation coefficients (p -value < 0.05) for parameters L, O, and D with respect to annual residential electricity consumption by number of dwellings in Andalusian municipalities (data source: SIMA-2019 [64]).

LOD Class	Parameter, x	ρ (x,y) y = Electric Consumption	ρ Size Effect	p-Value Significance
D	Average household size	0.55	BIG (+)	5.11×10^{-54}
D	Average No. of dwellings per residential building	0.51		1.19×10^{-45}
O	Average annual income per household	0.50		1.16×10^{-43}
D	Average artificial nighttime light glare (NTL)	0.46	MEDIUM (+)	5.96×10^{-36}
D	Average height of residential buildings	0.43		2.48×10^{-31}
L	Urban heat island, night–day difference	0.35		3.99×10^{-21}
L	Summer climatic severity (SCS)	0.20	SMALL (+)	2.84×10^{-07} *
O	Unemployment rate	0.03		5.04×10^{-1} **
D	Density of principal dwellings in urban area	−0.08	SMALL (−)	4.74×10^{-2} **
L	Global solar radiation	−0.12		2.07×10^{-3} *
D	Average floor area of dwellings	−0.16		3.40×10^{-5} *
L	Winter climatic severity (WCS)	−0.23		2.12×10^{-9} *
L	Elevation, height above sea level	−0.33	MEDIUM (−)	8.19×10^{-19}
D	Ratio of single-family buildings to total	−0.47		1.19×10^{-45}
O	Percentage of single-person households	−0.53	BIG (−)	3.69×10^{-50}
O	Percentage of population over 65 years old	−0.54		1.58×10^{-51}
O	Average population age	−0.55		1.01×10^{-54}
D	Average age of residential buildings	−0.56		5.30×10^{-55}

LOD classes: L, location; O, occupants; D, dwellings. ρ Size effect: BIG > 0.50 ; 0.3 $<$ MEDIUM < 0.5 ; 0.1 $<$ SMALL < 0.3 . Significance level: (*) low and (**) null.

It is worth noting that demographic parameters appeared to exert a greater influence than other factors, with climate showing a comparatively lower impact on household electricity consumption. Data on urban and building morphology had an intermediate weight in the influence. Although the unemployment rate is included in similar studies, in this case, null incidence and low significance were observed at this scale. Nevertheless, it should be noted that the computed electrical energy consumption value was not available in all localities, showing clear distortions in some cases. Consequently, after filtering, the final normalized SIMA dataset registered only 665 of the 785 possible records for Andalusia. Furthermore, the electricity consumption registered did not represent the entire final energy consumption of the residential sector in these municipalities. The electricity consumption considered has been distributed by local utilities companies operating in most of the municipalities in Andalusia, applying the ‘residential tariff’. This tariff also registers electricity consumption attributed to many small businesses and tertiary uses, resulting in deviations between what is expected and what is declared, as described in SPAHOUSEC I [76].

The irregular results observed with the higher prevalence of demographic data over climatic data in REC cannot and should not be extrapolated to general energy consumption. Climate variability across Andalusia affects the type of HVAC installation used in dwellings and, in turn, the consumption of a mixture of different energy sources, apart from electricity [77,78].

3.1.2. LOD Parameter Selection

According to the correlation analysis and additional strategic and methodological considerations detailed above, the parameters finally adopted at the regional–municipal scale for the LOD data model are shown in Table 2, with a statistical summary for Andalusia. The feature selection included the highest correlated parameters as well as others often

posited in the literature as linked to energy consumption. The complete LOD dataset for municipalities in Andalusia can be accessed in Appendix A.

Table 2. REC predictive parameters considered in the LOD data model for Andalusia: statistics.

Category	Par.	REC Predictive Parameter Description	Unit	Mean	Std. Dev
L, Location	l1	Winter climatic severity (WinCS)	-	0.69	0.23
	l2	Summer climatic severity (SumCS)	-	1.20	0.17
	l3	Annual global solar radiation	kWh/m ²	5.00	0.25
	l4	Elevation, height above sea level	m	503	333
	l5	Distance from the coastline	km	64	48
	l6	Maximum temperature in July	°C	32	2.30
O, Occupants	o1	Average population age	years	44.0	4.20
	o2	Percentage of population under 18 years old	%	15.0	4.30
	o3	Percentage of population over 65 years old	%	21.0	6.20
	o4	Average household size	Persons	2.4	0.28
	o5	Unemployment rate	%	24.0	5.30
	o6	Average annual income per household	€	21,500	3522
D, Dwellings, and urban environment	d1	Density of principal dwellings in urban area	dwellings/ha	32	17
	d2	Average artificial nighttime light glare (NTL)	nW/(cm ² sr)	225	517
	d3	Average height of residential buildings	floors	1.50	0.47
	d4	Average age of residential buildings	years	44	9.10
	d5	Average floor area of dwellings	m ²	130	19
	d6	Percent of single-family to total residential buildings	%	80.0	21
	d7	Average No. of dwellings per residential building	dwellings	1.40	0.87

3.2. Location, Occupant, and Dwelling Clustering

The application of clustering algorithms to the LOD dataset for each individual category resulted in a set of clusters, as analyzed below, as well as in the allocation of the potential REC process and discussion.

3.2.1. Location

Based on the statistical summary for each L, O, and D cluster, the potential REC ranking could be established for the individual categories among the clusters observed. Therefore, analysis was required prior to completing the PotREC assignment process of the L, O, and D categories.

The statistical summary of the characteristic parameters analyzed to perform each cluster according to the L, O, and D profiles is shown in Tables 3–5.

Table 3. Statistical summary and PotREC_L allocation of location clusters.

	Location	Unit	Average					Ranking					a *
	Parameter **		Location Cluster					Location Cluster					
			L0	L1	L2	L3	L4	L0	L1	L2	L3	L4	
l1	WinCS	-	1.12	0.77	0.49	0.41	0.77	5	4	2	1	3	1
l2	SumCS	-	1.03	1.10	1.37	1.14	1.37	1	2	5	3	4	1
l3	AGSolRad.	kWh/m ²	5.18	5.15	4.79	4.98	4.98	5	4	1	3	2	1
l4	Elevation	m	1013	668	156	171	517	5	4	1	2	3	1
l5	DCoastline	km	70	48	76	16	143	3	2	4	1	5	1
l6	TMaxJuly	°C	32	31	34	30	34	3	2	5	1	4	1
Location Score								22	18	18	11	21	
Location Potential REC, PotREC_L								5	3	2	1	4	

* a = Weighting coefficients for l₁ to l₆ for the location parameter vector in Equation (2). ** Description in Table 2.

Table 4. Statistical summary and PotREC_O allocation of occupant clusters.

	Occupant	Unit	Average					Ranking					b *
	Parameter **		Occupant Cluster					Occupant Cluster					
			O0	O1	O2	O3	O4	O0	O1	O2	O3	O4	
o1	AvePopAge #	years	49.70	40.67	46.08	42.79	49.45	-	-	-	-	-	0
o2	Pop < 18	%	7.50	18.88	13.63	16.96	10.14	1	5	3	4	2	1
o3	Pop > 65 #	years	25.50	15.38	23.33	18.74	28.19	-	-	-	-	-	0
o4	HousSize	Person	1.89	2.69	2.19	2.51	2.09	1	5	3	4	2	1
o5	Unemploy #	%	37.05	22.00	30.46	26.10	21.65	-	-	-	-	-	0
o6	HousAveInc	€	16,720	25,373	18,253	21,362	19,628	1	5	2	4	3	1
	HousElREC ***	kWh	2337	3724	2372	3370	2499	2	5	3	4	1	1
Occupant Score								5	20	11	16	8	
Occupant Potential REC, PotREC_L								1	5	3	4	2	

* b = Weighting coefficients for o₁ to o₆ for the occupant parameter vector in Equation (2). ** Description in Table 2.

*** Annual electrical consumption per dwelling analyzed in Section 3.1.1 (source: SIMA dataset). # Parameter not evaluated for PotREC_L estimation.

Table 5. Statistical summary and PotREC_D allocation for dwelling clusters.

	Dwellings	Unit	Average					Ranking					c *
	Parameter **		Dwelling Cluster					Dwelling Cluster					
			D0	D1	D2	D3	D4	D0	D1	D2	D3	D4	
d1	dwellings/ha	dwel/ha	31.70	17.30	23.26	21.91	18.46	1	4	2	3	5	1
d2	AveNTL	nW/cm ² sr	723	189	300	179	92	1	3	2	4	5	1
d3	AveHghRB	floors	2.10	1.48	1.78	1.93	1.08	1	4	3	2	5	1
d4	AveAgeRB	year	34.18	36.32	35.45	45.76	52.18	1	3	2	4	5	1
d5	DwAveFlAr	m ²	102.41	133.53	119.96	144.03	134.61	1	3	2	5	4	1
d6	SigFamRB	%	31.60	90.47	61.70	85.94	94.16	1	4	2	3	5	1
d7	Dw/Bld	dwellings	3.16	1.10	1.47	1.20	1.01	1	4	2	3	5	1
Dwelling Score								7	25	15	24	34	
Dwelling Potential REC, PotREC_L								1	4	2	3	5	

* c = Weighting coefficients for d₁ to d₇ for the dwelling parameter vector in Equation (2). ** Description in Table 2.

The map in Figure 6 presents the five clusters found. In keeping with their geo-statistical analysis, they were identified and named as “L0—Mountain climate”, “L1—Penibaetic and Aracena mountain ranges”, “L2—Guadalquivir Valley”, “L3—Coastal”, and “L4—Sierra Morena Range”.

Each location cluster found displayed a uniquely characteristic profile, as well as a clearly identifiable geographical distribution linked to well-known homogeneous geographical areas of Andalusia. Accordingly, cluster L0 represents the towns with the most severe winter climate, at an average altitude of 1013 m.a.s.l., in the provinces of Granada and Jaén. Clusters L1 and L4 showed certain similarities, including inland populations with average elevations of between 500 and 600 m.a.s.l. located in mountainous areas with a Mediterranean-continental climate, with mild to very cold winters and warm or hot summers, in the case of L4 (Sierra Morena). L2 is clearly distributed around the Betic depression or the inland Guadalquivir valley, displaying the highest mean July temperatures and summer climatic severity, with mild or temperate winters.

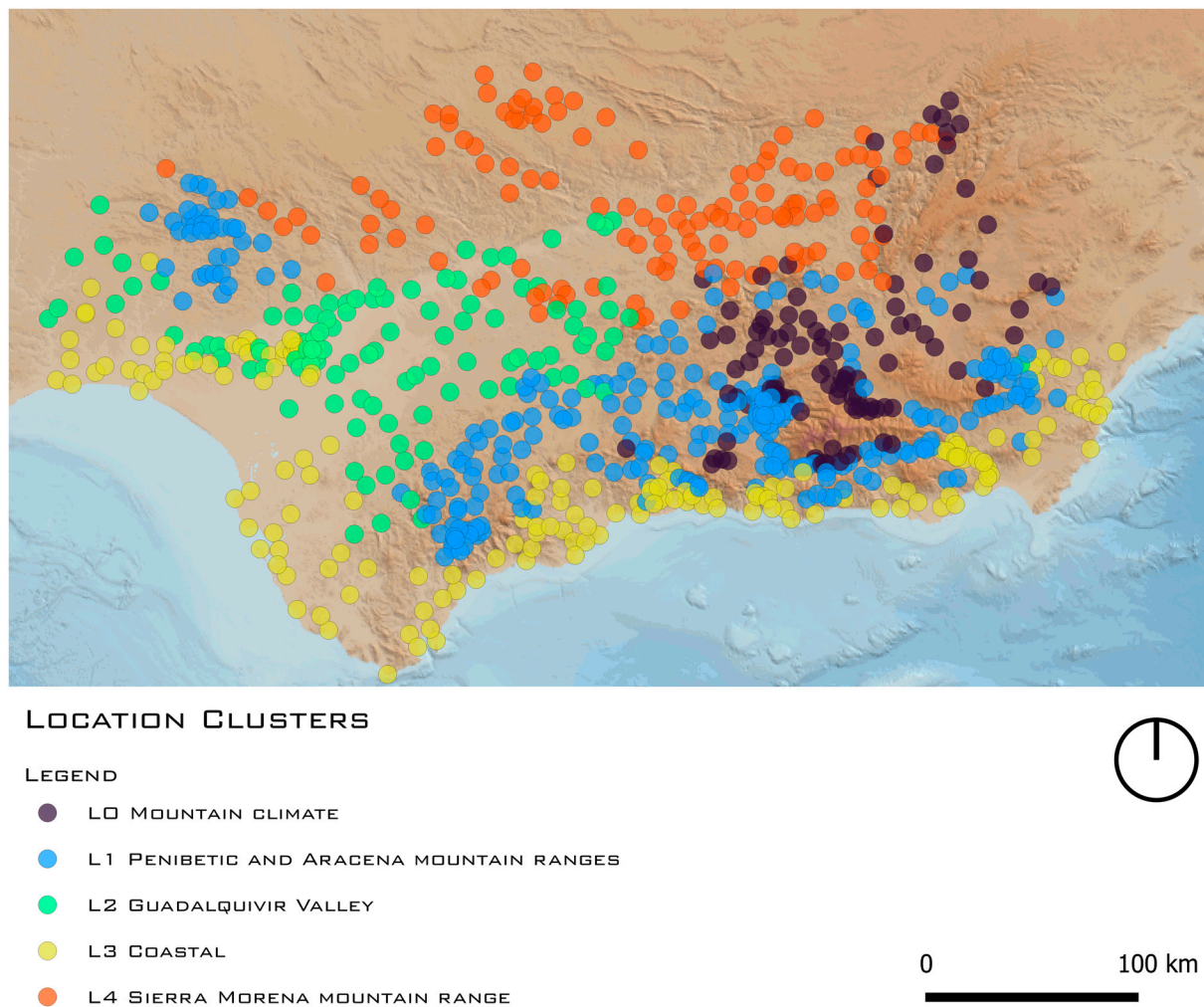


Figure 6. Location clusters. Distribution and description.

Finally, the L3 municipalities are distributed along the entire Andalusian coast, with mild winters and warm to hot summers and the lowest average maximum July temperatures of the dataset

The criteria applied when assigning PotREC_L corresponds to the combination of winter and summer climate severity and extreme temperatures, while elevation or distance from the coast are indicators of a more severe climate, especially in winter. In decreasing order, these localities can be classified as follows: the high mountain zone L0, the highlands L4 and L1, Guadalquivir valley L2, and finally, the coast L3. Consequently, the allocation was as expected, depending on the amount of heating consumption known for each area.

3.2.2. Occupants

The occupant clustering process resulted in four main clusters: “O1—Population in capitals and large municipalities”, “O2—Small populations in mountain areas”, “O3—Medium-sized towns and cities”, and “O4—Older population in rural areas”. The distribution and statistical summary of the characteristics of occupant-class clusters identified can be found in Table 4 and Figure 7.

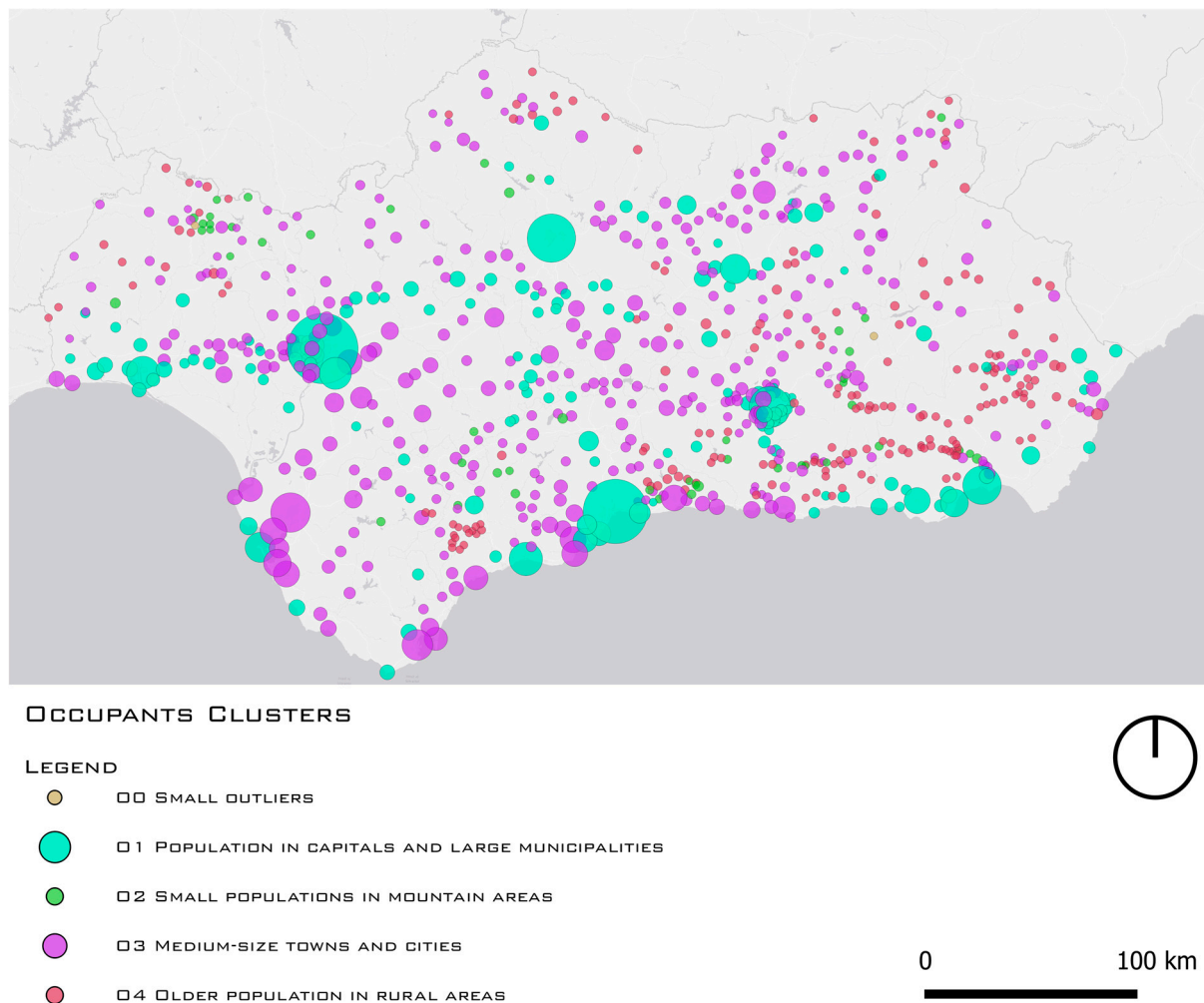


Figure 7. Occupant clusters. Distribution and description.

The profiles for clusters O1 and O3 combined demographic growth indicators (lower average age, higher rate of young people, a lower elderly population, and more people per household), whereas the reverse situation was observed in O2 and O4. An examination of economic indicators also showed coherence, since O1 and O3 had lower unemployment rates and higher incomes, unlike O2 and O4, although O4 combined lower incomes and unemployment rates.

Cluster O0, with only two municipalities, was considered an outlier of O4, and could in fact be included in the latter, despite its extreme combination of aging population, unemployment, and low incomes. It should be stressed that in the case of the PotREC_O assignment process, as expected, the resulting grade (1–5) assigned matched the exact order of O clusters in the ranking of persons per household (higher number of occupants commonly implies higher consumption per household).

3.2.3. Dwellings

Based on the residential building stock in Andalusian municipalities, the clustering process followed here enabled us to find five town clusters within this category, presented in Figure 8: “D0—Cities and large municipalities”, “D1—Low-density towns”, “D2—Municipalities in metropolitan areas”, “D3—Inland towns with large houses from the 1960s and 1970s”, and “D4—Small villages with old single-family houses”.

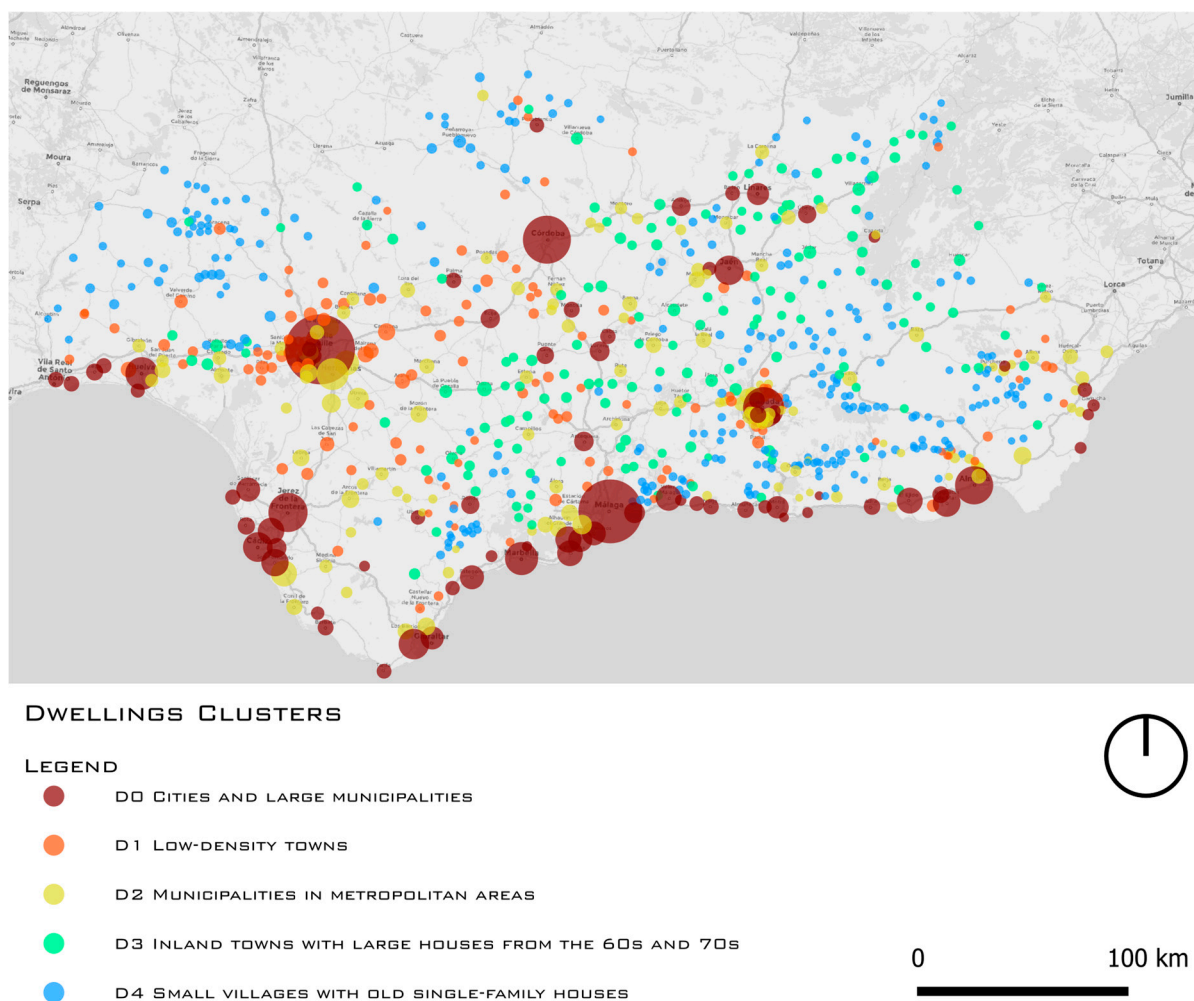


Figure 8. Distribution and description of dwelling clusters.

The results obtained, summarized in Table 5, show that cluster D0 corresponded to municipalities with a clear urban layout, characteristic of cities, including the highest dwelling density, the highest NTL brightness, and the highest share of high-rise multi-family buildings of the whole set.

Clusters D1 and D3 offered a very close match to their profile, although, as seen in Figure 8, location D1 is concentrated in western Andalusia and D3 in eastern Andalusia. Nevertheless, D3 presented higher percentages of multi-family buildings and average dwelling sizes, with higher numbers of constructions, which were, on average, noticeably older (by almost nine years) than those found in the western cluster. Cluster D2, located in high-population towns around the capital cities, presented the same profile in age and building typology as D0. The last cluster, D4, is located in small mountain villages, made up mostly of older single-family housing stock, with the highest average age of all. As predicted, the resulting PotREC_D assignment followed the same order as the average age of the dwellings of the individual clusters. Therefore, capital cities and metropolitan towns theoretically have the most efficient residential stock, while the buildings in small rural villages, which are mostly larger single-family houses, with older construction and fixtures, present the highest potential energy consumption.

3.3. LOD Cross-Clustering

Once the L, O. and D clusters for each municipality were obtained, an additional grouping process, named ‘cross-clustering’, as described in Section 2, was carried out. This process was used to identify any individuals simultaneously belonging to the same L, O,

and D clusters. It is considered essential in order to address the issue of potential multicollinearity arising from the clustering of L, O, and D factors. According to combinatory rules, taking 5 possible values with 3-by-3 repetition resulted in 75 possible combinations of L, O, and D clusters. However, only 72 different combinations were found when applying the clustering algorithm to the Andalusian dataset for these categorical parameters (L, O, and D categories). These combinations resulted in a set of 72 LOC cross-clusters detected in the final process. As 7 of the clusters found had only 1 member, the final number of LOD cross-clusters was reduced to 65.

This then allowed 99.10% of the municipalities, 99.79% of the main housing stock, and 99.82% of the population to be grouped into 1 of the 65 clusters detected.

A set of 10 relevant clusters has been identified, covering 21% of the total number of municipalities and representing 74% of the population, 73% of the dwellings, and 67% of the potential REC of the region. The rest of the clusters obtained, found mostly in rural populations, can be considered less significant, as their characteristics were less representative and differed greatly.

The distribution pattern of PotREC_LOD (Figure 5) showed the variability of this parameter and its coherence with the existing geographic and socioeconomic context throughout the region.

3.4. LOD Cross-Clustering Sample Application

An example of the application or implementation of the methodology is presented, identifying towns that should be prioritized in terms of energy intervention, for example, in retrofitting funds. Using the L, O, and D clusters and PotREC indicators obtained as starting points, LOD cross-clusters were identified, based on initial search criteria according to their respective L, O, and D characteristics. Therefore, a search process inverse to that of clustering was conducted, predefining the characteristics of the set to identify the locations of the municipalities meeting them. To optimize investment socially and technically, the search located the municipalities whose profile, expressed in Table 6, matched the combination of the most unfavorable conditions of vulnerability to energy poverty: extreme climate, inefficient housing stock, and a population with low purchasing power.

Table 6. Example of methodology implementation. Identification of clusters by search profiles characterized from PotREC L, O, and D indicators: at risk of energy vulnerability.

PotREC L, O, D	Clusters L, O, D	Cluster Description
		L (Location), O (Occupants), D (Dwellings)
PotREC_L = 5 (max.)	L0	L—Mountain climate
PotREC_O = 2 (min.)	O4	O—Older population in rural areas
PotREC_V = 5 (max.)	V4	D—Small villages with old single-family houses
LOD Cluster (5, 2, 5)	C65	Situation of energy vulnerability

According to this search profile, the LOD cluster of municipalities of the case was C65. LOD cluster C65 has a profile of energy vulnerability and includes a group of 39 small municipalities and villages located in mountain areas in eastern Andalusia.

3.5. Bottom-Up Data-Driven Model for Estimation of Provincial REC Based on Municipal LOD Parameters

This study validated the current information obtained for the set of towns in Andalusia. Here, we tested a new bottom-up statistical energy model [21] based on the LOD Potential REC of the municipalities for each Andalusian province and its number of dwellings. This new ‘Model 1’ was then compared with official REC consolidated statistics and two additional statistical models based in each province, dwelling ‘Model 2’ or population ‘Model 3’. The three models compared were, therefore, based on the weighted proportion

of regional REC to each province population, dwelling, or PotREC_LOD by the number of dwellings.

The mean absolute error (MAE), with respect to official statistics, for each model was obtained, as described in Table 7, showing the calculation of the regional REC distribution unitary coefficients of each model for the individual provinces.

Table 7. Determination of the mean absolute error (MAE) of the bottom-up models for estimating the provincial REC distribution.

	Province								Andalusia	Mean Absolute Error, MAE (%)
	Almeria	Cadiz	Cordoba	Granada	Huelva	Jaen	Malaga	Seville	Total Sum	
Consolidated Data										
2018 REC (ktoe) ¹	133.20	220.90	193.00	271.20	107.40	182.00	376.60	425.10	1909.60	
Share ²	0.0698	0.1157	0.1011	0.1420	0.0562	0.0953	0.1972	0.2226	1.0000	
Model 1 (PotREC_LOD)										
Dw·PotREC_LOD ³	1,861,271	2,977,533	2,635,020	3,391,391	1,567,658	2,616,150	4,361,031	6,009,855	25,419,909	
Share ²	0.0732	0.1171	0.1037	0.1334	0.0617	0.1029	0.1716	0.2364	1.0000	
Absolute error ⁴	0.0035	0.0015	0.0026	0.0086	0.0054	0.0076	0.0257	0.0138	0.0686	
Model 2 (No. of Dwellings)										
Dwellings	251,022	449,076	297,435	358,251	190,630	247,993	606,995	700,061	3,101,463	
Share ²	0.0809	0.1448	0.0959	0.1155	0.0615	0.0800	0.1957	0.2257	1.0000	
Absolute error ⁴	0.0112	0.0291	0.0052	0.0265	0.0052	0.0153	0.0015	0.0031	0.0972	
Model 3 (Population)										
Population	727,945	1,244,049	781,451	919,168	524,278	631,381	1,685,920	1,950,219	8,464,411	
Share ²	0.0860	0.1470	0.0923	0.1086	0.0619	0.0746	0.1992	0.2304	1	
Absolute error ⁴	0.0162	0.0313	0.0087	0.0334	0.0057	0.0207	0.0020	0.0078	0.1259	

¹ Source: REC-2018, Annual energy figures. *Consumo Energía Final Sector Residencial. Agencia Andaluza de la Energía* [79]. ² Province share with respect to total region, expressed in parts per unit. ³ Total sum of province towns (value of number of dwellings × PotREC_LOD). ⁴ Absolute error with respect to 2018 REC share.

The Model 1 MAE for estimating provincial REC was 0.86%, significantly lower than Model 2 (MAE 1.12%) and Model 3 (MAE 1.57%). This demonstrates the reliability of the application of this model based on the current information and knowledge obtained from the research process.

4. Conclusions

Typically, available studies and statistics on energy consumption in residential building stock in the region do not present enough detail and resolution to be applied to energy modeling at regional or urban scales. They often present significant deviations, specifically in the estimation of the energy consumption of air conditioning of households. Furthermore, real sub-regional figures of actual residential final energy consumption are not publicly available in most of the cases. An open data research approach helps enrich the analysis capacity and provide adequate resources for analysis. The methodology proposed here provides a predictive model for REC from regional to municipal scales. As this model is open and adaptable to different specific groupings with common characteristics (clusters), it can be replicated in other regions or territories with the proper settings. One of the strengths of this approach is its capacity to customize the influence of parameter categories on the model depending on the desired study approach or scope.

In this case, based on the selection of parameters from three categories: L (location), O (occupants), and D (dwellings), a clustering model of municipalities was provided and applied to the region of Andalusia. When applied to the case of this region, it resulted in a

total set of 65 clusters—or sets of municipalities—as detailed in the results annex. These were obtained from open datasets and had a high or medium correlation with the REC.

The clustering model was used to classify and identify the characteristic profile (LOD) in relation to REC, not only in terms of territorial distribution for individual municipalities, as in the case study, but also for metropolitan areas, associations of municipalities, counties, provinces, etc.

The method provided a bottom-up predictive model of REC at the regional–municipal scale, based on open data. Its application made it possible to obtain the potential REC (PotREC) of a city, metropolitan area, county, province, region, or any other type of grouping considered. When applied to the case study, a 0.86% mean absolute error (MAE) was obtained between the REC predictive model and actual historical data. This shows the degree of reliability of this model.

Among other actions, the implementation allowed the identification of the most relevant sets of municipalities in quantitative and qualitative terms on the potential REC. In the case study, the set of the first ten clusters included 21% of the municipalities, accounting for more than 73% of the population and dwellings of Andalusia, and 67% of the potential REC of the region. This predictive model could also be applied to identify clusters of municipalities in a situation of energy vulnerability or potentially highly suitable for energy rehabilitation policies, thus opening future lines of research of possible interest in the planning and management of resources at the regional–municipal scale.

The LOD clusters identified for the case study of Andalusia indicate future approaches detailing their common characteristics regarding potential REC and practical applications in planning regional- to town-scale energy strategies, helping stakeholders and policymakers to improve the performance and decarbonization of the residential building stock.

Future analysis of the correlation between the predictor parameters and the municipal REC recorded in forthcoming improved quality datasets is expected to contribute to optimizing the clustering processes, generating knowledge and a better understanding of their influence in consumption.

The future implementation of regional clustering using INSPIRE #250 m statistical cells or census units across a study region presents great potential as a territorial analysis tool, from regional to sub-urban scales with continuous resolution, regardless of the size of the municipality.

Author Contributions: Conceptualization, methodology, validation, formal analysis, investigation, resources, writing—review and editing, supervision, J.G.-L., J.J.S. and S.D.-A.; software, data curation, writing—original draft preparation, visualization, J.G.-L.; project administration and funding acquisition, J.J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data will be made available upon request.

Acknowledgments: The authors want to thank to the research group ‘USE-TEP 130 Architecture, Heritage, and Sustainability, Plan Andaluz de Investigación’ for their support. They also thank all open data initiatives held in Spain and Andalusia, which made accessing data for this research possible.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

AI	Artificial intelligence
DM	Data mining
EUI	Energy use intensity
GIS	Geographic information system
IECA	Institute of Statistics and Cartography of Andalusia
ESCO	Acronym for “Energy Services Company”

INSPIRE	Acronym for “Infrastructure for Spatial Information in Europe”
KDD	Knowledge discovery in databases
MAE	Mean absolute error
ML	Machine learning
PotREC	Potential residential energy consumption
REC	Residential energy consumption
UBEM	Urban building energy modeling

Appendix A. Dataset Links

Description	File Temporal Link
Complete L, O, and D Andalusia municipalities dataset	https://www.dropbox.com/scl/fi/cunc25fn7iuo5fe8z04rd/1-AND-LOD_DATASET-CLUSTERS.xlsx?dl=0&rlkey=ocmccg3pj6ic32h0bkph2i6el
Andalusia LOD Cross-Clusters dataset	https://www.dropbox.com/scl/fi/sq6wnn1vze9nx3vrblbto/2-AND-LOD_CROSS-CLUSTERS_AGGREGATED.xlsx?dl=0&rlkey=2tasfufi37in8otebouwlm5qp
GIS archive (Geopackage)	https://www.dropbox.com/scl/fi/cmwbrbnqty9c3dji9qc5q/ANDALUSIA_LOD_DATA-CLUSTER.gpkg?dl=0&rlkey=o6de8475sstpdovygkw30v9nh
SIMA dataset	https://www.dropbox.com/scl/fi/sia6uwdpba5bivxgtc97k/SIMA2019SelLOD.csv?dl=0&rlkey=g3cdy4q7ye56ej89302grph5p
SIMA correlation table	https://www.dropbox.com/scl/fi/bksj18jw108tjldyobcby/SIMA2019-CORRELATION-TABLE.csv?dl=0&rlkey=jmdxii7owkxycygyxn6tk8hz
SIMA correlation chart	https://www.dropbox.com/scl/fi/qpy1ojboq8k0wh5cwmdv1/CorrPlot.tiff?dl=0&rlkey=p18yp92spjci50gcoa173vsco

References

1. IDAE. Balance del Consumo de Energía Final. Available online: <https://www.idae.es/informacion-y-publicaciones/estudios-informes-y-estadisticas/estadisticas-y-balance-energetico> (accessed on 29 April 2023).
2. Balaras, C.A.; Dascalaki, E.G.; Patsioti, M.; Drousta, K.G.; Kontoyiannidis, S.; Cholewa, T. Carbon and Greenhouse Gas Emissions from Electricity Consumption in European Union Buildings. *Buildings* **2024**, *14*, 71. [\[CrossRef\]](#)
3. Eurostat (European Commission). *Manual for Statistics on Energy Consumption in Households, MESH*, 1st ed.; García Montes, J.P., Ed.; Eurostat-European Commission: Luxembourg, 2013; ISBN 978-92-79-33007-0.
4. Bednar, D.J.; Reames, T.G.; Keoleian, G.A. The Intersection of Energy and Justice: Modeling the Spatial, Racial/Ethnic and Socioeconomic Patterns of Urban Residential Heating Consumption and Efficiency in Detroit, Michigan. *Energy Build.* **2017**, *143*, 25–34. [\[CrossRef\]](#)
5. Reames, T.G. Targeting Energy Justice: Exploring Spatial, Racial/Ethnic and Socioeconomic Disparities in Urban Residential Heating Energy Efficiency. *Energy Policy* **2016**, *97*, 549–558. [\[CrossRef\]](#)
6. Arbulu, M.; Perez-Bezos, S.; Figueroa-Lopez, A.; Oregi, X. Opportunities and Barriers of Calibrating Residential Building Performance Simulation Models Using Monitored and Survey-Based Occupant Behavioural Data: A Case Study in Northern Spain. *Buildings* **2024**, *14*, 1911. [\[CrossRef\]](#)
7. Wang, S.; Sun, S.; Zhao, E.; Wang, S. Urban and Rural Differences with Regional Assessment of Household Energy Consumption in China. *Energy* **2021**, *232*, 121091. [\[CrossRef\]](#)
8. IDAE. *Consumos del Sector Residencial en España*; IDAE: Madrid, Spain, 2011.
9. IDAE. *SPAHOUSEC II: Análisis Estadístico del Consumo de Gas Natural en las Viviendas Principales con Calefacción Individual*; IDAE: Madrid, Spain, 2019.
10. IDAE. Estudio SPAHOUSEC III. Se Inicia la Recopilación de Datos de Consumo Energético de los Hogares. Available online: <https://www.idae.es/noticias/estudio-spa-housec-iii-se-inicia-la-recopilacion-de-datos-de-consumo-energetico-de-los> (accessed on 2 January 2022).
11. Berthou, T.; Duplessis, B.; Stabat, P.; Riviere, P.; Marchio, D. Urban Energy Models Validation in Data Scarcity Context: Case of the Electricity Consumption in the French Residential Sector. In Proceedings of the Building Simulation Conference Proceedings, Rome, Italy, 2–4 September 2019; Volume 5, pp. 3140–3147.
12. Müller, A.; Hummel, M.; Kranzl, L.; Fallahnejad, M.; Büchele, R. Open Source Data for Gross Floor Area and Heat Demand Density on the Hectare Level for EU 28. *Energies* **2019**, *12*, 4789. [\[CrossRef\]](#)
13. Domínguez-Amarillo, S.; Fernández-Agüera, J.; Peacock, A.; Acosta, I. Energy Related Practices in Mediterranean Low-Income Housing. *Build. Res. Inf.* **2019**, *48*, 34–52. [\[CrossRef\]](#)

14. García-López, J.; Sendra, J.J. Método Mixto Para la Determinación del Consumo de Climatización en Viviendas. In Proceedings of the 3rd International Congress on Sustainable Construction and Eco-Efficient Solutions, Seville, Spain, 27–29 March 2017; Mercader-Moyano, P., Ed.; Springer International Publishing: Seville, Spain, 2017; pp. 402–414.
15. Sendra, J.J.; Domínguez-Amarillo, S.; Bustamante, P.; León, A.L. Intervención Energética en el Sector Residencial del Sur de España: Retos Actuales. *Inf. Constr.* **2013**, *65*, 457–464. [\[CrossRef\]](#)
16. Gómez-Gil, M.; Sesana, M.M.; Salvalai, G.; Espinosa-Fernández, A.; López-Mesa, B. The Digital Building Logbook as a Gateway Linked to Existing National Data Sources: The Cases of Spain and Italy. *J. Build. Eng.* **2023**, *63*, 105461. [\[CrossRef\]](#)
17. Nastasi, B.; Manfren, M.; Noussan, M. Open Data and Energy Analytics. *Energies* **2020**, *13*, 5–7. [\[CrossRef\]](#)
18. Fremouw, M.; Bagaini, A.; De Pascali, P. Energy Potential Mapping: Open Data in Support of Urban Transition Planning. *Energies* **2020**, *13*, 1264. [\[CrossRef\]](#)
19. sEEnergies.eu. Pan-European Thermal Atlas 5.2. Available online: <https://euf.maps.arcgis.com/apps/webappviewer/index.html?id=8d51f3708ea54fb9b732ba0c94409133> (accessed on 5 December 2023).
20. Gomis, A.V. *Energía y Ciudades*, 1st ed.; de la Energía, C.E., Ed.; ENERCLUB: Madrid, Spain, 2017; ISBN 9788469737798.
21. Swan, L.G.; Ugursal, V.I. Modeling of End-Use Energy Consumption in the Residential Sector: A Review of Modeling Techniques. *Renew. Sustain. Energy Rev.* **2009**, *13*, 1819–1835. [\[CrossRef\]](#)
22. Zhao, H.X.; Magoulès, F. A Review on the Prediction of Building Energy Consumption. *Renew. Sustain. Energy Rev.* **2012**, *16*, 3586–3592. [\[CrossRef\]](#)
23. Lu, Y.; Chen, Q.; Yu, M.; Wu, Z.; Huang, C.; Fu, J.; Yu, Z.; Yao, J. Exploring Spatial and Environmental Heterogeneity Affecting Energy Consumption in Commercial Buildings Using Machine Learning. *Sustain. Cities Soc.* **2023**, *95*, 104586. [\[CrossRef\]](#)
24. De Jaeger, I.; Reynders, G.; Callebaut, C.; Saelens, D. A Building Clustering Approach for Urban Energy Simulations. *Energy Build.* **2020**, *208*, 109671. [\[CrossRef\]](#)
25. Seyedzadeh, S.; Glesk, I.; Roper, M. Machine Learning for Estimation of Building Energy Consumption and Performance: A Review. *Vis. Eng.* **2018**, *20*, 5. [\[CrossRef\]](#)
26. Wei, Y.; Zhang, X.; Shi, Y.; Xia, L.; Pan, S.; Wu, J.; Han, M.; Zhao, X. A Review of Data-Driven Approaches for Prediction and Classification of Building Energy Consumption. *Renew. Sustain. Energy Rev.* **2018**, *82*, 1027–1047. [\[CrossRef\]](#)
27. Amasyali, K.; El-Gohary, N.M. A Review of Data-Driven Building Energy Consumption Prediction Studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. [\[CrossRef\]](#)
28. Re Ceconi, F.; Khodabakhshian, A.; Rampini, L. Data-Driven Decision Support System for Building Stocks Energy Retrofit Policy. *J. Build. Eng.* **2022**, *54*, 104633. [\[CrossRef\]](#)
29. Reinhart, C.F.; Cerezo-Davila, C. Urban Building Energy Modeling—A Review of a Nascent Field. *Build. Environ.* **2016**, *97*, 196–202. [\[CrossRef\]](#)
30. Mastrucci, A.; Baume, O.; Stazi, F.; Leopold, U. Estimating Energy Savings for the Residential Building Stock of an Entire City: A GIS-Based Statistical Downscaling Approach Applied to Rotterdam. *Energy Build.* **2014**, *75*, 358–367. [\[CrossRef\]](#)
31. Nouvel, R.; Mastrucci, A.; Leopold, U.; Baume, O.; Coors, V.; Eicker, U. Combining GIS-Based Statistical and Engineering Urban Heat Consumption Models: Towards a New Framework for Multi-Scale Policy Support. *Energy Build.* **2015**, *107*, 204–212. [\[CrossRef\]](#)
32. Cerezo-Davila, C.; Reinhart, C.F.; Bemis, J.L. Modeling Boston: A Workflow for the Efficient Generation and Maintenance of Urban Building Energy Models from Existing Geospatial Datasets. *Energy* **2016**, *117*, 237–250. [\[CrossRef\]](#)
33. Caro-Martínez, R.; Sendra, J.J. Implementation of Urban Building Energy Modeling in Historic Districts. Seville as Case-Study. *Int. J. Sustain. Dev. Plan.* **2018**, *13*, 528–540. [\[CrossRef\]](#)
34. García-López, J.; Sendra, J.J.; Domínguez-Amarillo, S. Validating ‘GIS-UBEM’—A Residential Open Data-Driven Urban Building Energy Model. *Sustainability* **2024**, *16*, 2599. [\[CrossRef\]](#)
35. Buckley, N.; Mills, G.; Reinhart, C.; Berzolla, Z.M. Using Urban Building Energy Modelling (UBEM) to Support the New European Union’s Green Deal: Case Study of Dublin Ireland. *Energy Build.* **2021**, *247*, 111115. [\[CrossRef\]](#)
36. Buckley, N.; Mills, G.; Fealy, R. An Inventory of Buildings in Dublin City for Energy Management. *Ir. Geogr.* **2020**, *53*, 4–22. [\[CrossRef\]](#)
37. Monteiro, C.S.; Pina, A.; Cerezo-Davila, C.; Reinhart, C.; Ferrão, P.; Cerezo, C.; Reinhart, C.; Ferrão, P. The Use of Multi-Detail Building Archetypes in Urban Energy Modelling. In Proceedings of the 8th International Conference on Sustainability in Energy and Buildings, Turin, Italy, 11–13 September 2016; Elsevier Ltd.: Amsterdam, The Netherlands; Volume 111, pp. 817–825.
38. Ma, J.; Cheng, J.C.P. Estimation of the Building Energy Use Intensity in the Urban Scale by Integrating GIS and Big Data Technology. *Appl. Energy* **2016**, *183*, 182–192. [\[CrossRef\]](#)
39. Mutani, G.; Todeschi, V. Space Heating Models at Urban Scale for Buildings in the City of Turin (Italy). *Energy Procedia* **2017**, *122*, 841–846. [\[CrossRef\]](#)
40. Camporeale, P.E.; Mercader-Moyano, P. A GIS-Based Methodology to Increase Energy Flexibility in Building Cluster through Deep Renovation: A Neighborhood in Seville. *Energy Build.* **2021**, *231*, 110573. [\[CrossRef\]](#)
41. Martín-Consuegra, F.; de Frutos, F.; Oteiza, I.; Agustín, H.A. Use of Cadastral Data to Assess Urban Scale Building Energy Loss. Application to a Deprived Quarter in Madrid. *Energy Build.* **2018**, *171*, 50–63. [\[CrossRef\]](#)
42. Alonso, C.; Martín-Consuegra, F.; Oteiza, I.; De Frutos, F.; González-Cruz, E.; Cuerdo-Vilches, T.; Frutos, B.; Pérez, G.; Fernández-Agüera, J.; Domínguez-Amarillo, S. New Integrative Tool for Assessing Vulnerable Urban Areas. Refurbishment Model for

- Energy Self-Sufficient and Bio-Healthy Neighbourhoods. Madrid, Spain. HABITA-RES. *Front. Built. Environ.* **2023**, *9*, 113205. [CrossRef]
43. Rawal, R. Developing Urban Energy Models for Indian Cities: A Case Study of Ahmedabad—NZE. Available online: <https://nzeb.in/webinars/policy/developing-urban-energy-models-for-indian-cities-a-case-study-of-ahmedabad/> (accessed on 12 May 2023).
 44. Rit, M.; Girard, R.; Villot, J.; Thorel, M.; Abdelouadoud, Y. Calibration Method for an Open Source Model to Simulate Building Energy at Territorial Scale. *Energy Build.* **2023**, *293*, 113205. [CrossRef]
 45. Li, X.; Yao, R.; Yu, W.; Meng, X.; Liu, M.; Short, A.; Li, B. Low Carbon Heating and Cooling of Residential Buildings in Cities in the Hot Summer and Cold Winter Zone—A Bottom-up Engineering Stock Modeling Approach. *J. Clean. Prod.* **2019**, *220*, 271–288. [CrossRef]
 46. Aghamolaei, R.; Shamsi, M.H.; Tahsildoost, M.; O'Donnell, J. Review of District-Scale Energy Performance Analysis: Outlooks towards Holistic Urban Frameworks. *Sustain. Cities Soc.* **2018**, *41*, 252–264. [CrossRef]
 47. Andriessen, J.; Baker, M.; Cordasco, G.; De Donato, R.; Malandrino, D.; Palmieri, G.; Pardijs, M.; Petta, A.; Pirozzi, D.; Scarano, V.; et al. Increasing Public Value through Co-Creation of Open Knowledge. In Proceedings of the 2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG), 19–21 April 2017; IEEE: Quito, Ecuador, 2017; pp. 47–54.
 48. Mason, H.; Wiggins, C. A Taxonomy of Data Science. Available online: <http://www.dataists.com/2010/09/a-taxonomy-of-data-science/> (accessed on 26 December 2021).
 49. Laser, H.; Guhr, S.; Martenson, J.-H.; Gless, J.; Görner, J.; Amendt, D.; Schantze, B.; Gerbel, S. Data Science as a Service—Prototyping an Integrated and Consolidated IT Infrastructure Combining Enterprise Self-Service Platform and Reproducible Research. *Int. J. Adv. Softw.* **2020**, *13*, 104–115.
 50. Moreira, J.; Carvalho, A.; Horvath, T. *A General Introduction to Data Analytics*; John Wiley & Sons: Hoboken, NJ, USA, 2018; ISBN 1119296242.
 51. INE. *Censo de Población y Viviendas 2011. Edificios y Viviendas. Datos Provisionales*; INE: Madrid, Spain, 2013.
 52. Cuerdo-Vilches, T.; Navas-Martín, M.Á.; Oteiza, I. Behavior Patterns, Energy Consumption and Comfort during COVID-19 Lockdown Related to Home Features, Socioeconomic Factors and Energy Poverty in Madrid. *Sustainability* **2021**, *13*, 5949. [CrossRef]
 53. de Frutos, F.; Cuerdo-Vilches, T.; Alonso, C.; Martín-Consuegra, F.; Frutos, B.; Oteiza, I.; Navas-Martín, M.Á. Indoor Environmental Quality and Consumption Patterns before and during the COVID-19 Lockdown in Twelve Social Dwellings in Madrid, Spain. *Sustainability* **2021**, *13*, 7700. [CrossRef]
 54. IECA. Visor Tipologías Constructivas Catastro. Available online: https://www.ideandalucia.es/catalogo/inspire/srv/spa/catalog.search#/metadata/277ad216-53f9-45f5-b2c2-b7680d057365_200058_es (accessed on 27 July 2024).
 55. AEMET. *Atlas Climático Ibérico (Iberian Climate Atlas)*; MAMRM, AEMET and Instituto de Meteorología de Portugal: Madrid, Spain, 2011; ISBN 978-84-7837-079-5. [CrossRef]
 56. Corrado, V.; Ballarini, I.; Madrazo, L.; Nemirovskij, G. Data Structuring for the Ontological Modelling of Urban Energy Systems: The Experience of the SEMANCO Project. *Sustain. Cities Soc.* **2015**, *14*, 223–235. [CrossRef]
 57. European Commission. SEMANCO Semantic Tools for Carbon Reduction in Urban Planning. Available online: <http://www.semanco-project.eu/> (accessed on 6 November 2019).
 58. Tian, W.; Choudhary, R.; Augenbroe, G.; Lee, S.H. Importance Analysis and Meta-Model Construction with Correlated Variables in Evaluation of Thermal Performance of Campus Buildings. *Build. Environ.* **2015**, *92*, 61–74. [CrossRef]
 59. Mattinen, M.K.; Heljo, J.; Vihola, J.; Kurvinen, A.; Lehtoranta, S.; Nissinen, A. Modeling and Visualization of Residential Sector Energy Consumption and Greenhouse Gas Emissions. *J. Clean. Prod.* **2014**, *81*, 70–80. [CrossRef]
 60. Mora-García, R.T.; Martí-Ciriquian, P. Desagregación Poblacional a Partir de Datos Catastrales. In Proceedings of the XXIV Congreso de la Asociación de Geógrafos Españoles (AGE). Análisis Espacial y Representación Geográfica, Innovación y Aplicación, Zaragoza, Spain, 28–30 October 2015; Riva, J., Ed.; pp. 305–314.
 61. Li, X.; Yao, R.; Liu, M.; Costanzo, V.; Yu, W.; Wang, W.; Short, A.; Li, B. Developing Urban Residential Reference Buildings Using Clustering Analysis of Satellite Images. *Energy Build.* **2018**, *169*, 417–429. [CrossRef]
 62. Levin, N.; Duke, Y. High Spatial Resolution Night-Time Light Images for Demographic and Socio-Economic Studies. *Remote Sens. Environ.* **2012**, *119*, 1–10. [CrossRef]
 63. Duzgun, B.; Koksall, M.A.; Bayindir, R. Assessing Drivers of Residential Energy Consumption in Turkey: 2000–2018. *Energy Sustain. Dev.* **2022**, *70*, 371–386. [CrossRef]
 64. Instituto de Estadística y Cartografía de Andalucía (IECA). Sistema de Información Multiterritorial de Andalucía, SIMA. Available online: <https://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/index.htm> (accessed on 31 May 2023).
 65. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
 66. Liu, G.; Yang, J.; Hao, Y.; Zhang, Y. Big Data-Informed Energy Efficiency Assessment of China Industry Sectors Based on K-Means Clustering. *J. Clean. Prod.* **2018**, *183*, 304–314. [CrossRef]
 67. Xu, D.; Zhang, Q.; Zhou, D.; Yang, Y.; Wang, Y.; Rogora, A. Local Climate Zone in Xi'an City: A Novel Classification Approach Employing Spatial Indicators and Supervised Classification. *Buildings* **2023**, *13*, 2806. [CrossRef]

68. Syakur, M.A.; Khotimah, B.K.; Rochman, E.M.S.; Satoto, B.D. Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conf. Ser. Mater. Sci. Eng.* **2018**, *336*, 012017. [[CrossRef](#)]
69. Buttitta, G.; Turner, W.J.N.; Neu, O.; Finn, D.P. Development of Occupancy-Integrated Archetypes: Use of Data Mining Clustering Techniques to Embed Occupant Behaviour Profiles in Archetypes. *Energy Build.* **2019**, *198*, 84–99. [[CrossRef](#)]
70. Tardioli, G.; Kerrigan, R.; Oates, M.; O'Donnell, J.; Finn, D.P. Identification of Representative Buildings and Building Groups in Urban Datasets Using a Novel Pre-Processing, Classification, Clustering and Predictive Modelling Approach. *Build. Environ.* **2018**, *140*, 90–106. [[CrossRef](#)]
71. Zhao, B.; Han, W. Research on Measuring Methods and Influencing Factors of Spatial Damage Degree of Historic Sites: A Case Study of Three Ancient Cities in Shanxi, China. *Buildings* **2023**, *13*, 2957. [[CrossRef](#)]
72. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006; ISBN 978-0-387-31073-2.
73. Sharma, P. Distance Metrics | Understand Distance Metric in Machine Learning. Available online: <https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/> (accessed on 31 May 2023).
74. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Academic Press, Inc.: New York, NY, USA, 2013; ISBN 0-12-179060-6.
75. Hsu, D. Identifying Key Variables and Interactions in Statistical Models Ofbuilding Energy Consumption Using Regularization. *Energy* **2015**, *83*, 144–155. [[CrossRef](#)]
76. IDAE. *Informe Final SPAHOUSEC*; idea: Madrid, Spain, 2011.
77. García-López, J.; Sendra, J.J. Mixed Method for Determining the Air-Conditioning Consumption in Households. Application to Andalusia. In *Proceedings of the 3rd International Congress on Sustainable Construction and Eco-Efficient Solutions*, Seville, Spain, 27–29 March 2017; Mercader-Moyano, P., Ed.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 388–401.
78. IECA. Encuesta Social 2018. *Hogares y Medio Ambiente en Andalucía*. Available online: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/badea/informe/anual?CodOper=b3_2074&idNode=34478 (accessed on 27 March 2021).
79. AAE. Info-Energía, Sistema de Explotación de Información. Agencia Andaluza de la Energía. 2020. Available online: <http://www.agenciaandaluzadelaenergia.es/info-web/principalController>. (accessed on 31 July 2021).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.