Journal of Building Engineering 86 (2024) 108912

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

Identification of residential building typologies by applying clustering techniques to cadastral data

Alejandro Martínez-Rocamora [a,*], Pilar Díaz-Cuevas [b], Juan Camarillo-Naranjo [b], David Gálvez-Ruiz [c], Patricia González-Vallejo [d]

[a] ArDiTec Research Group. Departamento de Construcciones Arquitectónicas II, Universidad de Sevilla, Av. Reina Mercedes, 4-a, Sevilla, Spain
[b] Departamento de Geografía Física y Análisis Geográfico Regional, Universidad de Sevilla, Doña María de Padilla, S/n, Sevilla, Spain
[c] Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Tarfia, S/n, Sevilla, Spain
[d] ArDiTec Research Group. Departamento de Ingeniería Gráfica, Universidad de Sevilla, Av. Reina Mercedes, 4-a, Sevilla, Spain

ARTICLE INFO

Keywords:
Building typologies
Clustering
Cadastre
Benchmarking
Energy retrofitting
Urban-scale analysis

ABSTRACT

Building typologies are usually classified according to their shape, distribution and construction features depending on the time period they were built. As a result, a subjective classification arises which highly depends on the criteria used to differentiate buildings. When this analysis is carried out at the urban scale, data sets become bigger, making patterns difficult to uncover. Mistakes in deciding the variables to classify buildings lead to incorrect typologies and, consequently, wrong results. The aim of this study is to test a new methodology based on clustering techniques to identify typologies related to energy retrofitting, which would allow obtaining a more objective classification and better pattern recognition by reducing human intervention. To that end, a data set from the Spanish cadastre is used, with additional information to reflect the influence of existing standards on constructive solutions. By applying three clustering techniques (Ward's method, Partitioning Around Medoids, and a combination of both), new proposals of building typologies are obtained and discussed in comparison to traditional classifications. The results show that the Ward's method produces building typologies with significantly high quality metrics and meaningfulness. The agglomeration coefficient is 99.8%, which indicates that hardly another hierarchical method could generate a better clusters structure. Six clusters comprise 86% of the dwellings as most occupants do not declare retrofitting works, thus not being reflected in the cadastre database. This research provides a new classification method that can notably influence the estimation of costs, environmental impact and cost effectiveness of energy retrofitting actions at urban scale.

1. Introduction

The identification of building typologies becomes crucial for the application of cost and environmental impact estimation models, whether at urban scale or for a considerable amount of case studies that are geographically dispersed. In both situations, information about the dwellings must be obtained to classify them into typologies according to their characteristics [1]. Subsequently, this assignment of typologies allows generalizing bills of quantities to the set of buildings belonging to a same typology. Finally, through the bill of quantities, it is possible to apply economic and environmental impact models [2]. For this reason, the correct selection of

* Corresponding author.
E-mail address: rocamora@us.es (A. Martínez-Rocamora).

https://doi.org/10.1016/j.jobe.2024.108912
Received 30 October 2023; Received in revised form 9 February 2024; Accepted 22 February 2024
Available online 23 February 2024

quantities to apply and, therefore, the assigned typology, is paramount to obtain estimates closer to reality for the different buildings.

The classification of buildings into typologies has traditionally been carried out based on their formal and constructive characteristics, that is, by number of floors, floor area, and constructive solutions used to solve their foundations, structure or enclosure elements [3]. These classifications entail a manual analysis of the available data, and are usually decided subjectively based on the parameters of interest for the objective of the study [4], so the typologies obtained highly depend on decision-making and the criteria followed by the professional or organization that makes them. Depending on the number of available features in the dataset, a manual selection of key features for organizing data into clusters can become unfathomable. Moreover, this process can lead to a series of mistakes, such as using wrong or ineffective variables to classify data, thereby producing confusing typologies. Contrary to this approach, the use of machine learning techniques to analyse a data set of buildings in order to establish typologies according to similarities in their features' values could be an alternative to traditional methods that allows getting a more objective classification by removing the influence of humans. This kind of result can be obtained by using clustering algorithms, which is an unsupervised learning technique that tries to find groups of similar characteristics in a data set. Among the advantages that clustering techniques can provide compared to traditional methods are the following [5–7].

- Uncover hidden patterns that may not be apparent from inspecting raw data.
- Gain better understanding on the relationships between data points.
- Identify outliers and anomalies.

Table 1
Summary of key aspects of the studies included in the literature review.

Reference	Aim	Approach/Method	Variables	Results/Findings
Ali et al. [13]	Optimize urban scale energy retrofit decisions using data-driven approaches	Data collection Statistical analysis Data pre-procesing Outlier detection Feature selection	Building statistics Geometric data Non-geometric data Energy performance Retrofitting costs Reports from experts	Retrofit solutions and optimal features for residential buildings in Dublin (Ireland)
Dascalaki et al. [14]	Massive assessment of energy conservation measures	Manual	Construction period Type of building Climate zone Energy-related features	24 energy-related building typologies
Ghanbari et al. [15]	Identify architectural patterns	Clustering/Cosine similarity formula	Dimensions Size of plans	Two main typologies in Talesh (Iran)
Jiménez-Pulido et al. [16]	Classify façade typologies	Manual	Resistant layer Air chamber Thermal insulation Finishing	47 façade systems characterized to help future interventions
Mata et al. [17]	Detect building typologies in four European countries	Manual	Type of building Year of construction Heating system Climatic zone	Assign energy efficiency information and generalize to buildings of the same typology
Muringathuparambil et al. [4]	Define typologies of low-cost buildings to detect energy efficiency patterns	Manual	Building size Age Constructive solutions Building layout Energy system Electricity cost HVAC system Heating/cooling cost	8 typologies in Cape Town
Pistore et al. [18]	Identify significant retrofit areas and priorities of intervention	Wrapper Feature Selection Random Forest Hierarchical clustering K-medoids clustering	Configuration features (constructive solutions of thermal envelope and HVAC systems) Configuration features (exposed envelope surface, volume, number of floors, floor height)	Two clusters (+1 with outliers) from configuration features Two clusters (+1 with outliers) from intervention features
Rodrigues et al. [19]	Classify floor plan design proposals	Clustering/Ward's method - Euclidean distance	Point distance Turning function Grid-based Tangent distance	9 clusters with tangent distance being the best descriptor
Shan et al. [20]	Classify buildings and identify defining variables for energy efficiency	K-means clustering and Random Forest	Energy consumption Orientation Height to width rate of façades Perimeter to surface rate of building Exposed surface to building volume rate	Three energy-related building typologies in Wuhan (China), each with a set of key geometric variables
Wu et al. [21]	Identify neighbourhood typologies	K-means clustering	Local Climate Zone Urban descriptors	Four main typologies for planning climate adaptation interventions

- Increase the efficiency of the analysis by making it easier to make sense of large and complex data sets.
- Reduce dimensionality and complexity.
- Enable better decision-making and problem-solving.

Clustering algorithms have already been used on other occasions in the building engineering research field, although for other purposes, as in the case of Naganathan et al. [8], who applied them as a pre-processing step prior to supervised learning to assign percentages of energy loss reduction, thus forming a semi-supervised energy model. Similarly, Bienvenido-Huertas et al. [9] used the k-means algorithm to obtain a climatic classification of new and restored buildings in Andalusia, which offered energy demands in winter and summer for the different clusters that were better differentiated than those of the climatic zones according to the Spanish Technical Building Code. However, there are scarce examples of the application of clustering algorithms to classify building typologies, as it will be shown in the state-of-the-art.

The application of estimation models at urban or higher scale is considered a possible approach to tackle a problem recently highlighted by various authors: it is necessary to establish reference values or benchmarks with which each impact study can compare its results to correctly evaluate the analysed building or project [10]. This requires two conditions to be met: all the studies should use the same calculation methodology, and a large set of case studies should exist in which the same methodology has been applied, thus allowing to obtain benchmarks of results classified by typology [11]. Moreover, benchmarks would support policies with sustainability requirements for new buildings [12].

This study presents an applied methodology to carry out the first step of this solution, that is, the identification and assignment of building typologies to a considerable set of cases (i.e. buildings). It is important to clarify that the identified typologies must include buildings with a similar behaviour with respect to energy efficiency and energy retrofitting needs, as the final aim of this identification is to be able to explore and compare different retrofitting options and their consequences in terms of life cycle economic cost and environmental impact for a set of buildings of a specific typology. As the main novelty, this study is intended to use objective methods for such classification through the application of machine learning models for clustering. To demonstrate the functionality of this methodology, these algorithms are applied to a data set of real estate for residential use prepared for a municipality in Andalusia, which contains data obtained directly from the General Directorate of Cadastre (GDC), other indicators of interest prepared from this data, and several variables that characterize constructive solutions for various elements of the buildings' thermal envelope and installations. In the next section, a state-of-the-art is presented, including recent studies related to the classification of building typologies through different methods. In Section 3, the materials and methods used in this study are thoroughly described. The results are presented and analysed in Section 4. In Section 5, a discussion in comparison to the methodologies applied in similar studies is carried out, to finally draw conclusions in Section 6.

2. State-of-the-art

The classification of building typologies is traditionally organized according to the objectives of the study. In this section, previous studies on this research field are reviewed to identify the key aspects of the methods employed to classify buildings and their main results, which are summarized in Table 1 to illustrate the differences with respect to the present study.

Mata et al. [17] performed a segmentation process of the built stock in four European countries based on the type of building, year of construction, main heating system, and climatic zone. This allowed them to deduce and assign to the different archetypes characteristic values for the thermal envelope, internal loads, and other data of interest from the energy point of view, to finally be able to generalize the results to the rest of the buildings. Jiménez-Pulido et al. [16] proposed a classification system for typologies of façades in Spain by connecting information from the Technical Building Code, the Constructive Elements Catalogue, and other databases such as the Institute of Construction Technology of Catalonia (ITeC) and the CE3X software for energy efficiency certification [22]. In total, they identified and characterised 47 different façade typologies in their country, with a comprehensive description of each one that would allow a more efficient management of future interventions in existing buildings.

Muringathuparambil et al. [4] developed building typologies of low-cost buildings to detect energy efficiency patterns, obtaining eight representative types to classify buildings in Cape Town. These typologies were established according to design parameters, and energy efficiency related features were described in a subsequent step to be able to simulate them with DesignBuilder. Therefore, these features did not influence the initial establishment of building typologies. On an opposite approach, Dascalaki et al. [14] added energy-related characteristics from the TABULA project [23–25] to available data on the Hellenic building stock to find energy-related building typologies that would allow a massive assessment of energy conservation measures. Three specific construction periods influenced the characterization of these typologies: pre-1980, where buildings did not have thermal insulation; 1981–2000, being partially or insufficiently insulated; and buildings constructed after 2000, with proper envelope thermal insulation. This derived into 24 building types, combining these three construction periods, two types of buildings (single family and multifamily) and four climate zones.

Other interesting studies based their approaches on the analysis of floor plans to detect building typologies. Ghanbari et al. [15] identified architectural patterns in a sample of 150 rural houses in Talesh (Iran) based on the similarity of dimensions and size of plans. To that end, they applied a conversion method of floor plans into two-dimensional vectors, obtaining data sets that were subsequently analysed by searching for similarities in the distribution of spaces. From this study they concluded that two main typologies could describe the architectural patterns in this area: single-story houses and Telar houses, that is, houses with a covered porch on their first floor. Rodrigues et al. [19] used clustering techniques to help designers analyse and classify a total of 72 floor plan design proposals whose geometry was described using four different shape representation methods: point distance, turning function, grid-based, and

tangent distance. Concretely, they applied the Ward's method [26] and the Euclidean distance as dissimilarity measure, which classified observations into 9 different clusters. Finally, they concluded that the tangent distance descriptor better captured floor plan shapes and generated fewer outliers in the identified clusters, while each representation had its advantages and disadvantages.

Closer to the developments of the present study, several authors have presented approaches where machine learning techniques are applied to analyse a considerable amount of data for typology detection purposes. For instance, Ali et al. [13] presented a classification of buildings by archetypes or typologies based on their constructive characteristics through a complex flow of data processing, including collection steps, statistical analysis, pre-processing, outlier detection, and selection of variables, among others. Thus, they realized that it is of vital importance that the data set contains only information highly related to the objective of the identification of typologies, eliminating irrelevant or redundant data in this process of data set depuration. Shan et al. [20] obtained three clusters to classify buildings from Wuhan, China, according to their energy consumption, to subsequently carry out an internal analysis using Random Forest to identify the most defining variables. The orientation and the rate between the building's height and its width for the various orientations were identified as the most important variables for low and medium energy consumption buildings, while only orientation and that rate for the south orientation resulted paramount for high energy consumption buildings. Wu et al. [21] employed k-means clustering analysis to the Local Climate Zone framework applied to Amsterdam, London and Paris, whose data set was enriched with other urban descriptors, to identify four main neighbourhood typologies that could be used for planning climate adaptation interventions. Finally, Pistore et al. [18] combined Wrapper Feature Selection, Random Forests, and hierarchical and k-medoids clustering to identify the most significant retrofit areas and priorities of intervention in educational buildings in the Province of Treviso, Italy. Specifically, they used hierarchical clustering to obtain the optimal number of clusters in the data set and to identify outliers to be discarded. Then, they fed the k-medoids clustering with the selected medoids of the clusters from the previous step. From this process, they obtained two clusters when considering configuration features, and another two with intervention features, and one extra cluster of outliers in each case. Finally, these were combined into four clusters to prioritize the interventions in their various constructive elements.

3. Materials and methods

3.1. Description of the case study

The selected case study is the municipality of Osuna (Spain). Located in the east sector of Seville's province (37° 14 N, 5° 06′ W, 322 m. a.s.l.), 87 km from the provincial capital, Seville, 90 km from Cordoba and 130 km from Malaga, the municipality of Osuna corresponds to an average Andalusian city, a historical village originally founded by the Roman empire in 44–43 b. C. The Muslims called it Oxona, which derived into Osuna after the conquest by Fernando III of Castilla. During the 13th century, it became a strategic point

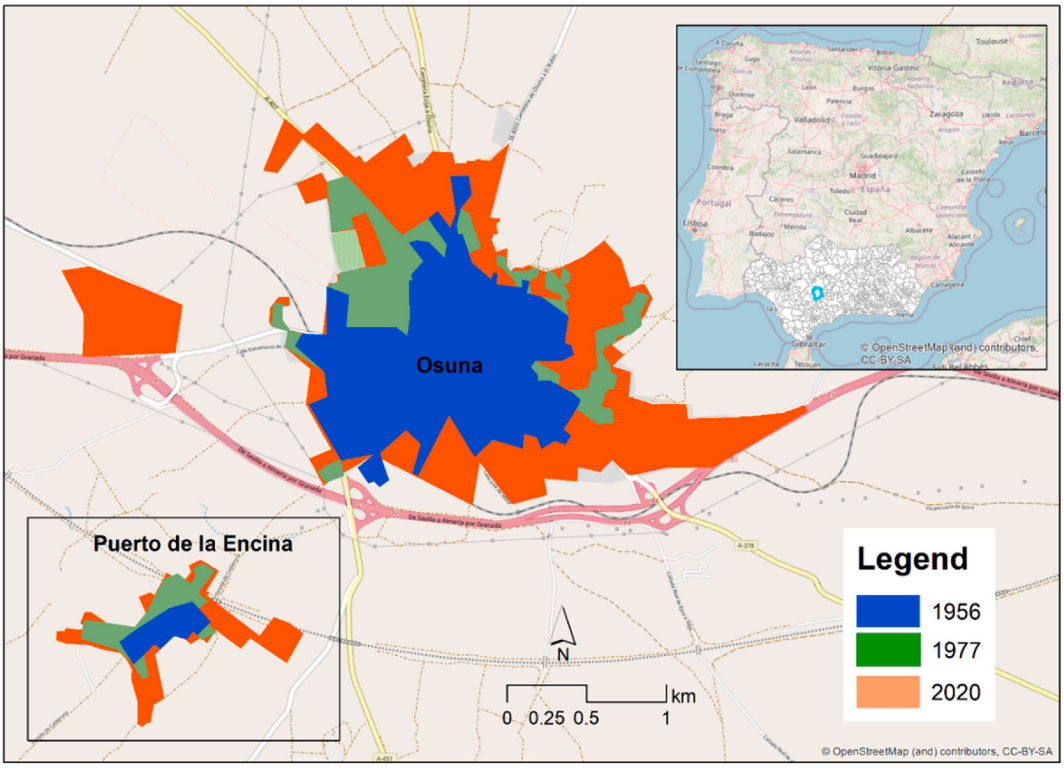

Fig. 1. Evolution of the built-up area of Osuna (1956–2020).

in the defence of the frontier with the Nasrid kingdom of Granada. Its population has remained stable since the middle of the 19th century with around 17,000 inhabitants, except in the period from the 1940s to the 1970s, when it increased to 23,000 inhabitants. The municipality is composed of two population centres, Osuna and Puerto de la Encina, the latter being less populated (218 inhabitants).

The evolution of the built-up area has been calculated by photo-interpreting and digitising it from the aerial orthophotographs of 1956, 1977 and 2020. Fig. 1 shows the location of the municipality of Osuna, as well as the evolution of the built-up area, which has increased by 2.5 km^2, with 1.6 km^2 in 1956, 2.14 km^2 in 1977 and, finally, 4.1 km^2 in 2020 (latest aerial orthophotography available at the time this work was carried out).

3.2. Spanish cadastre database structure

The General Directorate of Cadastre in Spain organizes its information through several structured text file formats. Concretely, a user can download SHP (Shapefile) and CAT (for Catastro, Cadastre in Spanish) files for a municipality that contain, respectively, graphical and non-graphical information [27]. The latter comprises surfaces, use, year of construction, among others. These two formats are linked through primary keys that allow representing non-graphical information in a map at parcel level, and their structure is described in separate manuals that establish every information field and their corresponding length [28,29]. This information can be downloaded through the official website of the GDC and is permanently updated, which makes the use of this source of information relatively frequent for different scientific research related to the subject of this work [30,31].

The data of a municipality is divided into several types of registers containing the values for a series of features corresponding to parcels, constructive units, constructions, real estate, distributions of common elements, and cropland. The proposed approach consists of identifying residential building typologies through the real estate they contain. Thus, only records for parcels, constructions and real estate need to be considered in this study. Constructive units are excluded because they do not provide useful information for this purpose.

As shown in Fig. 2, in a parcel there can be one or more buildings, each one of them comprising one or more real estate. Additionally, a real estate can be made up of one or more constructions. For example, on many occasions there is a construction for each floor of a single-family house, as can be seen in the 2-story building of street A in Fig. 2. Parcels are the minimum graphic unit to represent data on the map, since buildings are not identified in the GDC database. For this reason, the organization of real estate by buildings within the same parcel is proposed as an intermediate objective of this study. To that end, it will be necessary to work with the fields that define the address: type of street, street name, first number, first letter, second number, second letter, kilometre, block, stair, floor and door. The floor and door fields are specific for real estate as they identify, for instance, an apartment in a building, such as those in the 4-story buildings in both streets of Fig. 2. Therefore, a building address goes all the way to the stair field, with the stair and block fields being often used interchangeably for the same purpose.

3.3. Clustering algorithms: parameters and basic concepts

Clustering algorithms are unsupervised learning techniques that consist of studying unlabelled data sets and trying to detect groups of similar observations according to the values of some of their features, organizing data in a way that similarities within a cluster are

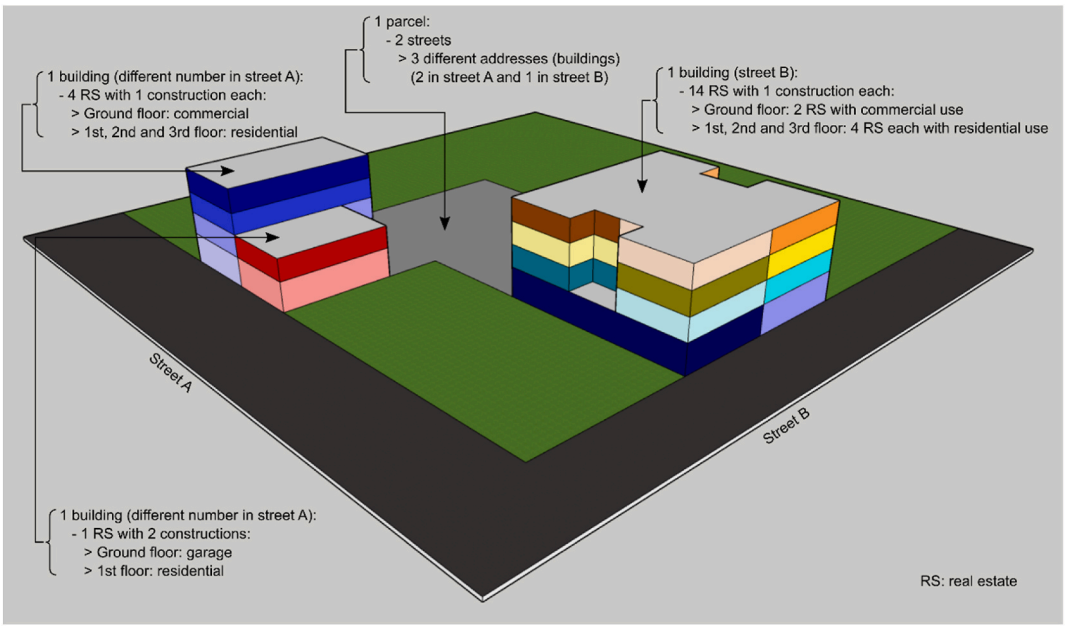

Fig. 2. Example of hierarchical organization of Cadastre entities for three buildings in a same parcel.

maximized while differences between clusters are minimized [32]. The main clustering algorithm categories include partition-based algorithms (k-means or PAM), hierarchical algorithms (agglomerative and divisive clustering using minimum, maximum, average, centroid or Ward linkage strategies), and density-based algorithms (topological-density based like DBSCAN or probability-density based EM-GMM). Each have their own strengths based on the aim of the research and the nature of the variables used [33], but clustering algorithms provide advantages for analysing unlabelled data through discovering hidden patterns [5], reducing dimensionality [6], identifying outliers [7], and not requiring predefined labels compared to supervised classification methods [33], which help reveal inherent structures within complex datasets.

3.3.1. Data matrix and clustering

The process begins with an $n \times m$ matrix (where n is the number of observations and m is the number of variables) or $n \times n$ matrix whose elements represent the distances or similarities between the n observations. Therefore, each observation is represented by an m-dimensional vector with the values of that observation in the different variables.

Let Ξ be the set of n observations, $(x_1,\ldots,x_n)$, where $x_j = (x_{j1},\ldots,x_{jm}), j = 1,\ldots,n$. The objective is to find a partition of the Ξ into "c" regions or clusters: C_1, …, C_c, such that:

$$\begin{aligned} \bigcup_{i=1}^{c} C_i &= \Xi \\ C_i \cap C_l &= \varnothing \quad i \neq l \end{aligned}$$

In matrix form, what is obtained is a matrix that provides the values of the variables for each observation:

$$X = \begin{matrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ x_{j1} & \cdots & x_{ji} & \cdots & x_{jm} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{ni} & \cdots & x_{nm} \end{matrix}$$

The j-th row of the matrix X contains the values of each variable for the j-th observation, while the i-th column shows the values belonging to the i-th variable across all observations in the data set.

From this matrix, it is necessary to establish the choice of the measure of association to be used to determine the similarity of the observations. This measure is defined as a distance between the elements of the data matrix.

3.3.2. Distance matrix

Given the existence of both qualitative and quantitative variables in our data set, it is necessary to apply the clustering methods on a pre-calculated distance matrix, instead of applying it directly to the original data set. To that end, the Gower's distance matrix is used, which accepts both types of variables for each observation. This disqualifies some of the abovementioned methods, as they do not accept a distance matrix as input.

The generalized definition of distance between two observations (real estates) in feature space i and j given by Gower [34] results in the dissimilarity defined as:

$$d_{ij} = \frac{\sum_{l=1}^{m} \delta_{ij}^{(l)} d_{ij}^{(l)}}{\sum_{l=1}^{m} \delta_{ij}^{(l)}} \in [0,1]$$

where, if x_{il} and x_{jl} are values for observations i and j on variable (l):

$d_{ij}^{(l)}$ = contribution of variable (l) to d_{ij}, which depends on its type:

- if (l) is binary or nominal: $d_{ij}^{(l)} = 0$ if $x_{il} = x_{jl}$, and $d_{ij}^{(l)} = 1$ otherwise
- if (l) is a scaled interval variable (continuous or cardinal discrete): $d_{ij}^{(l)} = \frac{|x_{il} - x_{jl}|}{\max_h x_{hl} - \min_h x_{hl}}$.
- if (l) is ordinal or proportional scale: for each observation i in each variable l the ranks r_{il} are calculated, and from them, $z_{il} = \frac{r_{il}-1}{\max_h r_{hl}-1}$ is treated as a scaled interval variable (continuous or cardinal discrete).

$\delta_{ij}^{(l)}$ = weight of each variable (l):

- $\delta_{ij}^{(l)} = 0$ if values of observations i and/or j for variable (l) are missing values, values of observations i and j for variable (l) are zero and variable (l) is an asymmetric binary variable.
- $\delta_{ij}^{(l)} = 1$ otherwise.

Once the dissimilarity metrics have been defined, it is necessary to choose the clustering technique to be used. As previously introduced, the need to use the distance matrix as input data in cluster techniques is a restrictive factor when identifying applicable clustering methods. Although distance matrices can be considered data matrices whose variables are now the distances of each

observation to the rest, this dramatically hinders interpretability in some techniques, such as those based on probabilistic density, like Expectation-Maximization clustering using Gaussian Mixture Models (EM-GMM). On the other hand, other density-based methods, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [35] or MeanShift, require the assignment of topological parameter values that assume a uniform density of observations throughout the data set, which is not desirable in this analysis. However, hierarchical methods, by dealing with distances between individuals in each iteration, and the Partitioning Around Medoids (PAM) method [33], become good candidates as they work directly through distances provided in the Gower distance matrix. Therefore, three clustering techniques are selected for this study: the Ward's method [26] or minimum variance as hierarchical clustering to generate compact clusters, the PAM density-based method, and a combination of these two where PAM is fed with the number of clusters and the medoids detected by the Ward's method (hereafter PAM-Ward). These last two methods do not generate a hierarchical structure (dendrogram) that allows studying how the union of elements evolves regarding their likeness.

3.3.3. Ward's clustering method

The Ward's method [26] or minimum variance is a hierarchical clustering technique. The goal of these methods is to merge or split clusters based on a distance or similarity measure. There are two types of hierarchical methods: agglomerative and divisive. Agglomerative methods begin with as many clusters as data points and merge them until there is only one cluster left. Divisive methods start with one cluster that contains all the data and split it until there is one cluster per data point. Divisive methods are harder but faster if the number of clusters is known beforehand, which is not our case.

All clustering methods use a distance or similarity matrix to decide how to form the clusters (Gower distance matrix in our case). Ward's criterion measures the information loss from clustering by the sum of squared deviations of each point from the cluster mean, or centroid. The optimal clustering minimizes this sum. The agglomerative version of Ward's method follows this rule: at each step, merge the two clusters that have the smallest increase in the sum of squared deviations from their new centroid.

All hierarchical methods can be expressed through a single recurrent formula (Lance–Williams formula) of four parameters, such that, for the different values of these, the different distances between clusters at each stage are generated [36]. According to this formula, if two clusters $C_1,$ and C_2 join in a new cluster $(C_1 \cup C_2),$ the dissimilarity between their union and a new cluster $C_3,$ is given by:

$$d((C_1 \cup C_2), C_3) = \alpha_1 * d(C_1, C_3) + \alpha_2 * d(C_2, C_3) + \beta * d(C_1, C_2) + \gamma * |d(C_1, C_3) - d(C_2, C_3)|$$

Ward method uses the formula with

$$\alpha_i = \frac{n_i + n_3}{n_1 + n_2 + n_3}; \beta = \frac{-n_3}{n_1 + n_2 + n_3}; \gamma = 0$$

Where n_1, n_2, n_3 are the number of elements in C_1, C_2, C_3.

3.3.4. PAM clustering method

PAM (Partitioning Around Medoids) is a specific partitioning clustering algorithm similar to K-Means. While K-means considers centroids (the mean position of all the points in a cluster, calculated as the component-wise average that may not necessarily correspond to any real data point) to determine the recurring partitions, PAM tries to find representative objects in each cluster, called medoids, that best characterize the cluster. These medoids are actual data points of clusters that have the least average dissimilarity to all the other data points in the cluster. Unlike centroids, medoids must be existing data objects rather than an average point. The PAM clustering method is developed in Ref. [31] and follows the following sequence.

1. It starts with an initial set of medoids, either selected randomly or through a heuristic method. Each data point is assigned to the cluster whose medoid is closest to that point.
2. Next it determines if swapping any currently assigned medoid with a non-medoid data point would reduce the total distance between points and medoids. This is the cost function.
3. It iteratively swaps medoids to try and minimize this cost function until there is no cost decrease from further swaps. This produces a locally optimal configuration of medoids and clusters.

PAM is well-suited for cases in which intuitive interpretations are needed since medoids are real examples from the data. It also is more robust to outliers as medoids are real data points not affected by outliers as centroids.

3.3.5. Clustering evaluation

The results from these clustering methods are evaluated through the most commonly used grouping quality metric, the Silhouette coefficient [37]. Moreover, this metric is also appropriate in order to identify critical observations in each generated cluster as the Silhouette coefficient can be calculated individually. In cluster analysis, the Silhouette coefficient measures the similarity of an observation to its assigned cluster compared to its similarity to the neighbouring clusters. Its value ranges from -1 to 1, where a value closer to 1 indicates that the observation is appropriately assigned to its cluster, and a value closer to -1 indicates that the observation is more similar to the neighbouring cluster than to its assigned cluster or can be considered an outlier [7]. In other words, the observation might be an outlier or an anomaly, and it does not fit well with the other observations in its assigned cluster, being adequate to be considered as a possible error. From the maximization of the average of these coefficients, the optimum number of clusters is obtained. In the case of the hierarchical clustering, its quality can also be measured with the agglomeration index or cophenetic correlation (AC), which describes how the dendrogram generated by the hierarchical cluster captures the information of

the original data set.

3.4. Research methodology

To achieve the main aim of this research, a sequential data processing method is followed as described in Fig. 3. This methodology can be divided into three main stages: data preprocessing, input feature selection, and implementation and performance evaluation. These last two actions are carried out in two steps, as the evaluation of results from the first execution allows obtaining feedback to improve the data set and, consequently, the results obtained.

3.4.1. Data preprocessing

In this stage, batch data for the municipality of Osuna, available for the year 2022, is first downloaded from the GDC website [27] in a CAT file. Based on the specification of the CAT format [28], a relational data model is designed and implemented in PostgreSQL/PostGIS object-relational database system by programming an SQL query algorithm that performs a structured reading (namely parsing) of the CAT file and dumps all the data of the different types of records to a PostgreSQL/PostGIS database. This spatial database management system optimises the query, exploitation, and update possibilities, as well as the different exploitation procedures of the spatial database created.

Within this system, a new package of SQL queries is developed that allows the creation of a real estate table with the original data of interest for the detection of typologies, as well as additional indicators obtained from the existing data. The parameters obtained for real estate, either directly or by calculation, are specified in Table 2 accompanied by a brief explanation of their meaning and the method for obtaining them.

With respect to the last set of variables in Table 2, the constructive solutions for those elements are identified by alphanumeric codes that represent the specific element, the type of dwelling, the construction period and, finally, the renovation period in case the element has undergone a renovation. The construction and renovation periods defined are based on the evolution of energy efficiency regulations for buildings in Spain, which determine significant milestones such as the requirement of thermal insulation in the building's envelope, improvements in the performance of windows through the use of double glazing or thermal-bridge break technologies, as well as the integration of renewable energy sources and enhancements in Heating, Ventilation and Air Conditioning (HVAC) and Domestic Hot Water (DHW) systems. More specifically, the MV regulations of 1957 [38] did not consider the thermal performance of buildings. Thermal insulation became mandatory with the NBE-CT79 regulation on thermal conditions in buildings [39]. In 2006, the Technical Building Code (CTE by its Spanish acronym) was approved [40], including a specific document on Energy Savings in Buildings (DB-HE), which had subsequent updates in 2013 and 2019. All this information is captured in the TABULA project [23,24], thus establishing the following construction and renovation periods: before 1936, 1937–1959, 1960–1979, 1980–2006, 2007–2013, 2014–2019, and after 2020.

After building this table, a post-processing stage begins aimed at removing wrong registers and to correct failures in the organization algorithm. In this stage, the constructions with an "OD" value (part of 'tODos', 'all' in Spanish, meaning that all the constructions belong to the same owner, which is usual in cadastral records of old dwellings) in the *floor* field (96 in total for this municipality), real estates with an empty *street_name* field, which impedes their organization by building through their address (377), a series of wrong real estate registers that caused interferences in the calculations (20–25), and real estates without any 'dwelling' use, are removed. Finally, it is necessary to apply some corrections to real estates taken by the algorithm as belonging to multi-family buildings while they are actually single-family terraced houses within a same parcel, only differentiated by their *door* field instead of *stair*. In total, 7150 registers of residential real estates are preserved as the data set to which the clustering algorithms are to be applied. Once these errors are corrected, the resulting table is exported as the definitive data set. All subsequent statistical and cluster related analysis is conducted using R [41], and RStudio [42].

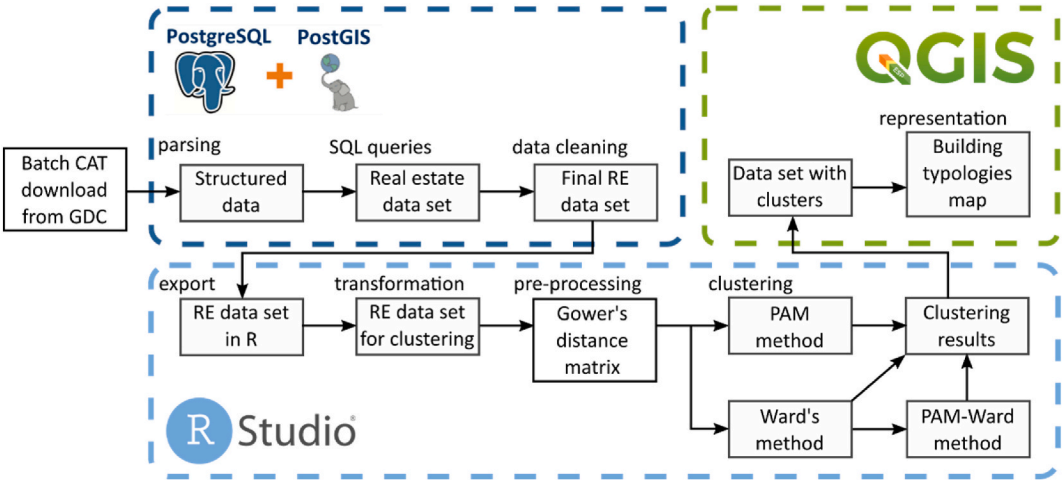

Fig. 3. Data processing flowchart.

Table 2
Variables and indicators for real estates and their definitions.

Variable/Indicator	Definition
cad_ref	Cadastral reference of the real estate.
RS_cod	Code of the real estate within its parcel.
cad_par	Code of the cadastral parcel.
parcel_surf	Surface of the parcel.
parcel_building	Number of its building within the parcel, obtained by numbering the different addresses within a same *cad_ref*.
total_parcel_build	Total number of building in the parcel, obtained as the maximum of *parcel_building* with a same *cad_par*.
coord_X	Coordinate X of the parcel.
coord_Y	Coordinate Y of the parcel.
municipality	Municipality to which the parcel belongs.
pop_municipality	Number of inhabitants of the municipality.
st_type, st_name, number, letter, sec_number, sec_letter, km, block, stair, floor, door	Variables for the definition of the real estate's address.
yr_constr	Year of construction of the real estate.
yr_last_reform	Year of the last reform to the real estate, obtained as the most recent year of reform of its constructions.
reform_type	Type of the last reform made to a construction of the real estate (R: integral rehabilitation, O: complete reform; E: intermediate reform, I: minimum reform).
conserv_status	Conservation status of the real estate, obtained from INSPIRE service and generalized to all the buildings within a same parcel.
RS_builtup_surf	Built-up surface of the real estate.
coef_property	Property coefficient of the real estate within its building.
Use	Use (D: dwelling)
RS_max_floor	Intermediate variable to obtain the number of floors above ground of the building. Calculated as the maximum *floor* of all the constructions of a real estate.
RS_min_floor	Intermediate variable to obtain the number of floors below ground of the building. Calculated as the minimum *floor* of all the constructions of a real estate.
floors_above_gr_build	Number of floors above ground of the building. Obtained as the maximum floor of all the real estates of the same building plus 1 (floor 0 is the first one above ground).
floors_below_gr_build	Number of floors below ground of the building. Obtained as the minimum floor of all the real estates of the same building, in absolute value.
building_type	Type of building (S: single-family, M: multi-family).
dw_per_floor	Intermediate variable for the number of dwellings per floor of the building. Number of dwellings in the same *floor* of the real estate.
max_dw_per_floor	Number of dwellings per floor of the building. Obtained as the maximum of *dw_per_floor* in the same building.
built_surf_per_floor	Intermediate variable for the built-up area per floor of the building. Built-up area of dwellings in the same *floor* of the real estate.
max_built_surf_per_floor	Built-up area of dwellings per floor of the building. Obtained as the maximum of *built_surf_per_floor* in the same building.
RS_built_surf_above_gr	Intermediate variable for the built-up area above ground of the building.
built_surf_above_gr_build	Built-up surface above ground of the building. Sum of the *RS_built_surf_above_gr* of the building.
RS_built_surf_below_gr	Intermediate variable for the built-up area below ground of the building.
built_surf_below_gr_build	Built-up surface below ground of the building. Sum of the *RS_built_surf_below_gr* of the building.
perc_resid_use	Percentage of residential use of the building, obtained from the use of its real estates and their respective built-up areas over the total of the building.
roofing, façade, sec_façade, windows, floor, heating, cooling, DHW	Variables to define the type of constructive solutions, generated combining the element, *building_type*, *yr_constr*, *yr_last_reform*, *reform_type*, and the evolution of solutions of the thermal envelope and installations according to project TABULA.

3.4.2. Input feature selection

Prior to applying clustering techniques, it is essential to prepare the data set adequately. One of the fundamental prerequisites for these algorithms is that all features must have a valid value, and no observation should contain any missing values. Furthermore, the clustering process requires a uniform structure of both qualitative and quantitative variables to avoid any potential errors that may arise by mixing variables of different nature. In this process, variables and indicators that should not participate in the clustering process are removed. Concretely, from those specified in Table 2, *cad_ref* is only used as unique primary key for each real estate, and those in grey are removed whether because they are intermediate variables to obtain others or because they are not of interest for the identification of residential building typologies for this study's aim. In the case under study, *municipality* and *pop_municipality* are removed as they present the same value for all the observations. However, if this methodology was applied to a data set from various municipalities, *pop_municipality* should be preserved.

Subsequently, all the empty values in the *reform_type* field are replaced by a new category named 'NoRef' indicating that no reform has been carried out in that real estate. The other main modification consists of creating a new variable named *YWOR* (years without reform) replacing *yr_last_reform*. Thus, *YWOR* represents the number of years passed since the last reform, thereby improving the harmonization of its values as a great number of similar values when there has not been any reform could be interpreted as a category within a quantitative variable.

3.4.3. Implementation and performance evaluation

As mentioned before, given that there are qualitative and quantitative variables in the data set, it is necessary to pre-calculate a Gower's distance matrix. Once this matrix has been obtained, it is used as input to the selected clustering methods, *i.e.* Ward, PAM-Ward and PAM. The results from each algorithm are evaluated using a clustering quality metric, the Silhouette coefficient, and the agglomeration index (AC), as explained in the previous subsection.

A first iteration of the algorithms and the analysis of the various clusters and their characteristics allows detecting some incoherences in the data set originated in the GDC database, as well as variables that could be removed since their value can be inferred from others. Regarding the incoherences or errors, real estates whose year of last reform according to the GDC database is previous to their year of construction are detected (33 in total), and others for which a complete reform or even an integral rehabilitation is recorded in the same year of construction, something improbable as these types of reforms imply modifying the façade and roofing systems of the building. In the second case, 104 registers have been found and, by visually inspecting their façades through online images, this information does not match the year of construction according to the database. Therefore, it is assumed that probably the actual year of construction was overwritten by mistake in the GDC database with the same value of the year of reform, thus making it impossible to recover it and obtain correct values for these registers. This could mean that, in the first case, the *yr_last_reform* and *yr_constr* values were somehow swop in the GDC database, but this cannot be assumed with certainty. As the resulting data loss is considered insignificant (1.9%), in order to make the necessary corrections, these registers that generate small clusters only containing outliers, are also removed. In this step, 137 observations are removed from the data set, leaving 7013 in total. Then, a second iteration of the clustering algorithms is applied to the new corrected data set to obtain the final results. Finally, a geographic information system, in this case QGIS, is employed to represent the results of the cluster analysis. As explained in the introduction, in forthcoming research, the assigned clusters will help to generalize energy efficiency analyses in order to explore alternatives of future improvements of their thermal envelope and installations.

4. Results

From the application of the clustering algorithms in two iterations, the results shown in Table 3 arise. As it can be observed, the number of identified clusters in the first iteration is higher than that of the second iteration as the removal of incoherent data allowed eliminating their corresponding clusters of outliers. Moreover, the *yr_constr* and *YWOR* variables are finally discarded from the process as the periods of construction and reform can be inferred from the codification applied to the constructive solutions variables and, in fact, these variables have a significant influence on the nature of the identified clusters. Quality metrics, as well as the number of observations with higher probability of a cluster assignment error, become better in the second iteration for every method. In the case of the Ward's method, the agglomeration coefficient obtained is 99.8%, which indicates that hardly another hierarchical method applied on the same data set could generate a better clusters structure. PAM-Ward method does not improve the results from Ward's method, so its cluster proposal can be ignored. PAM method obtains its best results with 16 clusters, almost half the quantity of clusters from Ward's method, which implies reorganizing and merging clusters. Despite this increases the number of possible errors (negative Silhouette coefficient for an observation), PAM method preserves a good mean Silhouette coefficient.

In Table 4, the 28 clusters identified by the Ward's method in the second iteration are described through the values of their most defining variables. As it can be seen, these variables are the type of building and the construction and reform periods, which are precisely those conditioning the values of the constructive solutions variables. Despite the number of floors was expected to be a decisive variable in the formation of clusters, this was not the case in Osuna. From the experts viewpoint, the ground truth in this aspect is that, depending on the study, its main aim and the data set, the number of floors is taken into account as an important variable or not. For example, for Shan et al. [20], the number of floors was one of the most defining variables. In contrast, the classification from the TABULA project [43] does not consider the number of floors among the defining variables for their typologies, thus obtaining four different typologies in Spain: single-family house, terraced house, multi-family house, and apartment block. In view of the results for Osuna, the meaningfulness of this might be explained by two main reasons: first, the uniformity of building heights in towns like Osuna, where the maximum number of floors is restricted in most of its area by their urban planning ordinances; and second, the low influence of the number of floors in other variables of the data set. Perhaps if the existence of an elevator in the building had been specified in an additional variable, this would get reflected in a partition of each cluster of multi-family buildings in two different typologies depending on the number of floors in which the elevator becomes mandatory.

With respect to the clusters description, there are six clusters containing the vast majority of dwellings in the data set (86.5%), four

Table 3
Results from the clustering algorithms expressed through indicators.

Iteration	No. Variables	Indicator	Ward	PAM-Ward	PAM
1	22	No. clusters	33	33	17
		Mean Silhouette coefficient	0.928	0.803	0.889
		No. possible errors	67	31	59
		% possible errors	0.937	0.434	1.021
2	20	No. clusters	28	28	16
		Mean Silhouette coefficient	0.933	0.819	0.906
		No. possible errors	53	32	73
		% possible errors	0.756	0.456	1.021

Table 4
Definition of clusters detected by Ward's method.

Cluster	Type of building	Floors above ground	Floors below ground	Construction period	Type of reform	Reform period	No. cases	No. pos. errors
1	S	1–3	0–1	2007–2013	–	–	810	0
2	S	1–3	0–1	2020-Now.	–	–	35	0
3	S	1–4	0–1	1980–2006	–	–	2016	0
4	S	1–4	0–1	1960–1979	–	–	706	0
5	S	1–4	0–1	1900–1936	O/E	1960–1979	1664	0
6	S	1–3	0–1	1937–1959	–	–	140	3
7	S	1–3	0–1	2014–2019	–	–	61	0
8	S	1–3	0	1900–1936	R/O/E	2014–2019	27	0
9	M	1–4	0–1	1980–2006	–	–	392	0
10	S	2–3	0–1	1920	R/O/E	2020-Act.	14	0
11	M	2–6	0–1	1960–1979	–	–	479	0
12	S	1–3	0	1960–1979	O/E	2014–2019	13	0
13	M	1–4	0	1920	O/E	1960–1979	96	0
14	S	1–3	0–1	1960–1979	O/E	1980–2006	27	0
15	S	1–3	0	1675–2013	-/R/O/E/I	1940–2021	57	57
16	S	1–3	0	1900–1936	–	–	27	9
17	S	1–3	0–1	1900–1936	R/O/E	1980–2006	150	0
18	S	1–4	0	1900–1936	O/E	2007–2013	67	0
19	S	2–3	0	1960–1979	R/O/E	2007–2013	19	0
20	M	2–3	0–1	2007–2013	–	–	71	0
21	M	2–3	0–1	1920	O/E	1980–2006	23	0
22	S	1–3	0	1980–2006	O/E/I	2007–2013	30	4
23	S	1–3	0–1	1980–2006	E	1980–2006	17	0
24	S	1–3	0	1960–1979	I	1980–2013	16	0
25	S	1–2	0	1937–1959	O/E	1980–2006	12	0
26	S	2–5	0–1	1980–2006	O/E	2014–2019	13	0
27	M	2–4	0	1920–2019	-/O/E	1960–2019	20	20
28	M	2–3	0	1960–1979	E/I	1980–2006	11	0

Type of building - > S: single-family; M: multi-family.
Type of reform - > R: integral rehabilitation; O: complete reform; E: intermediate reform; I: minimum reform.

of them comprising single-family houses (1, 3, 4 and 5), and two of them dwellings in multi-family buildings (9 and 11). More specifically, these clusters include single-family houses built in successive periods from 1960 to 2013 and the oldest houses in the municipality, built in the 1900–1936 period and reformed in the 1960–1979 period. Clusters 9 and 11 include multi-family buildings built from 1960 to 2006 which have no reform recorded in the database. At this point, it is worth noting that the recording of reforms entirely depends on the registration of construction licenses, which are not always processed by homeowners, thereby not being included in the GDC database. It is also to be mentioned that clusters 15 and 27 (highlighted in grey) are entirely composed of outliers with very low or negative Silhouette coefficients. Due to the heterogeneity of their observations, these clusters cannot be easily described and therefore must be discarded from this study.

On the other hand, as it was mentioned before, the results from the PAM method (see Table 5) show a lower number of clusters at

Table 5
Definition of clusters detected by PAM method.

Cluster	Type of building	Floors above ground	Floors below ground	Construction period	Type of reform	Reform period	No. cases	No. pos. errors
1	S	1–3	0–1	2007–2013	–	–	813	3
2	S	1–3	0–1	2020-Now.	–	–	38	3
3	S	1–4	0–1	1980–2006	–	–	2022	6
4	S	1–4	0–1	1960–1979	–	–	716	10
5	S	1–4	0–1	1900–1936	O/E	1960–1979	1675	8
6	S	1–3	0–1	1937–1959	N/I	1980–2013	152	12
7	S	1–3	0–1	2014–2019	–	–	62	1
8	S	1–3	0	1920	R/O/E	2014–2019	49	22
9	M	1–4	0–1	1980–2006	–	–	397	5
10	S	1–3	0–1	1900–1936	R/O/E/I	1980–2006	164	8
11	M	2–6	0–1	1960–1979	–	–	494	15
12	S	1–3	0–1	1900–1979	O/E/I	1966-Act.	67	41
13	M	1–4	0	1920	O/E	1960–2006	130	11
14	S	1–5	0–1	1900–2013	R/O/E/I	1940-Act.	68	52
15	S	1–4	0–1	−2013	-/R/O/E/I	1940-Act.	95	28
16	M	2–3	0–1	2007–2013	–	–	71	0

Type of building - > S: single-family; M: multi-family.
Type of reform - > R: integral rehabilitation; O: complete reform; E: intermediate reform; I: minimum reform.

the expense of merging some of the Ward's method smaller clusters, thus increasing the number of possible errors identified in each cluster. Moreover, in this case four clusters of outliers are generated (8, 12, 14 and 15, highlighted in grey). However, the same main clusters are obtained with a similar number of dwellings, comprising 87.2% of the total observations. Thus, the selection of the most adequate algorithm relies on the decision between a higher clustering quality with few assignment errors but higher segmentation or, on the contrary, less segmentation with more possible errors with clusters that generalize better. In view of these results and considering that the aim of this clustering process is to characterize building typologies by their energy efficiency profile, the Ward's method would be a good option as it identifies them more precisely regarding their construction and reform periods, which directly influence the constructive solutions for the thermal envelope and installations.

Finally, a cartographic representation of the main clusters identified by the Ward's method is shown in Fig. 4. Since graphical representation is restricted to a cadastral parcel level, not being able to represent values for each building, the value retrieved from the most observed cluster among the real states within each parcel (modal cluster) has been used as the parcel's value. In this map, a certain uniformity in the distribution of clusters can be observed, as contiguous areas tend to be urbanized with similar types of buildings in the same time period, which also gives us an idea of the urban development process in Osuna over the years. The only typology that includes reformed buildings is number 5, which comprises 1664 cases. These cases correspond to the oldest single-family houses (SFH) in the municipality, dating back to the period between 1900 and 1936. They are primarily located in the historic centre and underwent renovations between 1960 and 1979. Buildings from the same time period (1960–1979) are also found in cluster 4, mainly situated on the outskirts with some remnants in the historic centre. Cluster 3, which encompasses the largest number of cases (2016), is distributed concentrically from the historic centre towards the outskirts of the nucleus. These buildings were built between 1980 and 2006. The most recent buildings are represented in cluster 1 and are primarily located in the southern part of the historic centre, on the outskirts. These buildings correspond to the area that has experienced the most significant growth in recent years. The remaining clusters with the highest number of individuals (9 and 11) correspond to multi-family houses (MFH) -or buildings-which are mostly distributed throughout the historic centre without a clear spatial pattern. It is worth noting that the majority of the parcels belonging to the remaining typologies (in grey), have been retrofitted, as can be seen in their respective clusters in Table 4.

Also the contours of the orthophotographs previously shown in Fig. 1 have been added to Fig. 4, as they allow an additional analysis about the degree of renovation of the different urban areas. As can be observed in Table 6, inside the contour of the orthophotography taken in 1956, 45% of the parcels have been built after the date of the flight, which means whether they were originally empty in the year the orthophotography was taken and built afterwards or they were demolished and re-built in later years. 6% out of that 45% and 98% out of the remaining 55% of the parcels have undergone rehabilitation works. This high rate of renovation is logical since this contour comprises the oldest buildings in Osuna. Regarding the orthophotography from 1977, 28% of the parcels were built or demolished and re-built after the date of the flight, with a small number of them being retrofitted afterwards as these can be considered relatively new constructions. Almost 100% of the remaining parcels, which were built between 1956 and 1977, have undergone

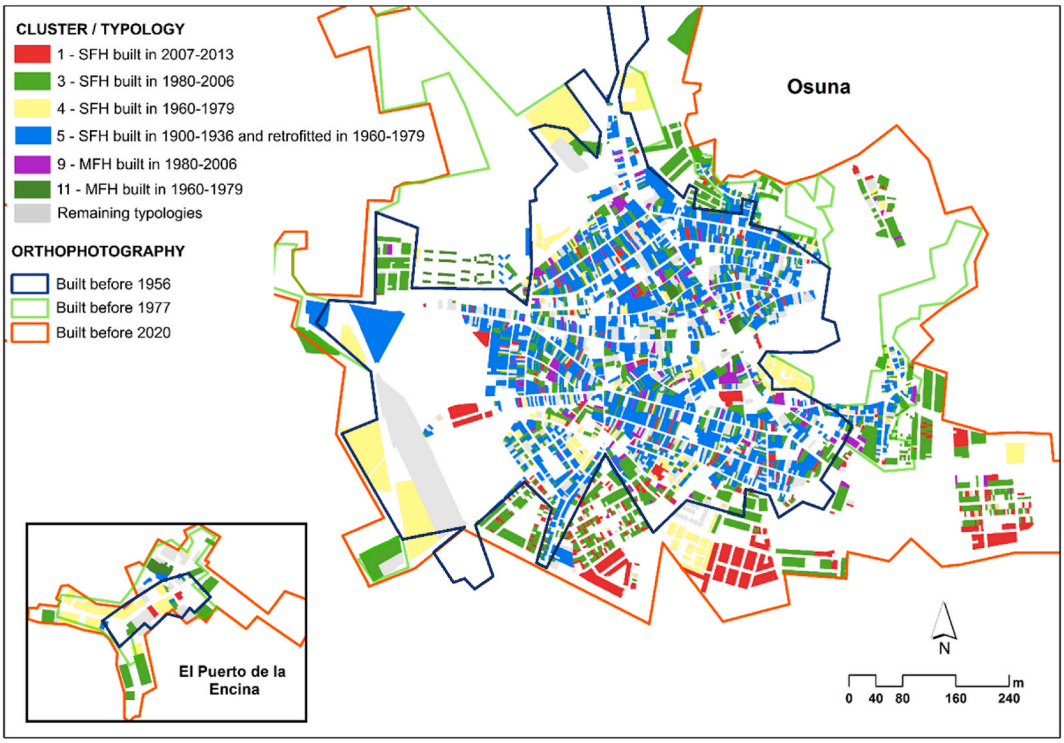

Fig. 4. Cartographic representation of the identified clusters in the municipality of Osuna.

Table 6
Renovation degree of the urban areas identified through orthophotographs.

OP year	Parcels with any residential RS	Construction after OP	Construction after OP, and retrofitted	Retrofitted after OP
1956	3259	1480	87	1743
1977	526	148	5	377
2020	1714	11	0	64
Total	**5499**	**1639**	**92**	**2184**

RS: real estate
OP: Orthophotography.

renovations after 1977. Finally, the orthophotography from 2020 identified 1714 additional parcels containing real estates with residential use built between 1977 and 2020. Only 11 (0.6%) have been built or demolished and re-built after 2020, and 64 (3.7%) have been retrofitted after the orthophotography year.

In total, 29.8% of the parcels have been built or re-built, and another 39.7% have been retrofitted, after their corresponding orthophotography was taken. This means that the information gathered from the Cadastre and treated automatically through the proposed methodology allows obtaining updated data for 70% of the parcels over what could be inferred from the available orthophotographs. Moreover, this information can be automatically updated without extra costs with each update of the GDC database. As a consequence, this allows generalizing building features through the cluster organization to obtain better estimates of renovation costs. The most frequent renovation period was 1960–1979, before the Spanish Technical Code of Construction was in force, which means Osuna will need to put a great effort and make significant renovation actions in the forthcoming years to comply with the requirements established by the European Union. Consequently, the final aim of the present study gains importance, as this methodology is oriented to estimate renovation costs at the urban scale by simplifying the calculations through the generalization that the obtained clusters allow.

5. Discussion

In this section, the methods applied in the present study are discussed in comparison to those from similar research, as results have been found to depend on the data set, thus making results singular for each study. Regarding the use of PostgreSQL/PostGIS to address the Cadastre database, there are several previous studies applying similar techniques. However, in this research some interesting approach differences have been proposed and put into practice. For example, Noguero Hernández et al. [44] used Cadastre data and a PostgreSQL/PostGIS spatial database management system to georeference and spatialise residential redevelopment actions at plot level, while Pérez-Alcántara et al. [45] designed a methodology based on PostgreSQL/PostGIS for the generation and visualization of housing indicators (GRID) through cadastral data. In addition, Pérez-Alcántara et al. [46] used this data integration methodology for the analysis and characterisation of residential space in the Andalusian coastline and its spatial representation in GRID format. Cadastral information and the design and implementation of a relational data model through PostgreSQL/PostGIS have also been used for the identification of housing for tourism purposes in the city of Seville [47] or to analyse urban land uses in Alcalá de Henares [48]. Mora García and Marti Ciriquian [30] merged data from census sections with information from the GDC to disaggregate population data at the cadastral parcel and building level, in an exercise that is similar to the first steps described in the methodology of the present study. With these data, they were able to geographically represent indicators such as the number of houses and inhabitants per hectare on density maps.

In this work, the authors go a step further and surpass the analyses carried out so far, at the real estate or parcel level, identifying and characterizing the buildings by processing the postal addresses of the real estates. In addition, unlike other papers that focus on real estates with dominant residential use, this study specifically selects those with at least one ‘dwelling’ use. This inclusion ensures that a significant number of dwellings are not left out of the analysis, as it is usual in towns like Osuna to have a garage in ground floor whose surface can be higher than that of the dwelling upstairs.

When comparing to other studies focused on defining building typologies, two different approaches have been detected. While Mata et al. [17] and Dascalaki et al. [14] considered energy-related characteristics among the data to classify buildings into typologies, Muringathuparambil et al. [4] included these in a subsequent step, thus not influencing the typology identification process. In the present study, this last option was discarded as, in order to obtain useful building typologies, the detection process should be influenced by features related to the main aim of the study, as stated by Ali et al. [13]. Obtaining additional energy-related features such as orientation, thermal envelope shape factors, or the energy efficiency label in order to improve the identification of typologies was not feasible due to the complexity of the connection between the data set and the shape of their geographical representation. This was a limitation of the present study because, despite the authors managed to organize the real estate by buildings, their geographical identification in the map is not available in the data set as the smallest geographical entity is the parcel.

Finally, it was interesting to notice that several studies have applied, in some point of their methodology, similar steps than those in this research, but then derived into a different path. For example, Rodrigues et al. [19] applied the Ward’s method and Euclidean distance instead of Gower’s to floor plan descriptors; Shan et al. [20] used clustering as a first step to subsequently identify important variables related to building design within each cluster through Random Forest; and Pistore et al. [18] employed hierarchical clustering to discard outliers, and then used the medoids obtained from the previous method to feed a *k*-medoids clustering (PAM). From this experience, the authors consider that the identified typologies were useful for the objective of finding groups of buildings with similar energy-related features, thus allowing to generalize results and explore the possibilities for a simplified urban-scale analysis,

always assuming the estimation errors that might arise from this simplification.

6. Conclusions

The main aim of the study, consisting of identifying building typologies by applying clustering techniques to a big real estate data set, was achieved obtaining good results according to the clustering quality metrics evaluated and to the analysis of their meaningfulness according to the final purpose of the study and to the ground truth of building typologies in the analysed municipality. The use of clustering algorithms allowed obtaining such typologies with an objective method; however, it was noticed that the variables defining the constructive solutions used for each element of the thermal envelope and installations somehow forced the formation of clusters. Despite that, further analysis on what would have happened if those variables had been omitted is not considered of interest as, for the purpose of this research, focused on the energy efficiency profile of building typologies, it was crucial to consider these-variables as they are required for generalizing the thermal characteristics of the medoid of each cluster to the rest of observations within it. Regarding this generalization, forthcoming research should focus on how to refine the assignment of results to other members of the cluster by adapting them according to their characteristics.

During the first stage of the study, in which data from the GDC database had to be extracted and organized, it was found that a translation of some of the address fields was necessary, as well as the debugging of wrong data contained in their database. The methodology employed in this study focuses on using relational spatial databases for analysis, setting it apart from alternative approaches that heavily rely on geographic information systems. This strategic choice not only simplifies the data updating process but also provides the added benefit of transferring the code to various research domains through scripts. Such an approach guarantees smooth integration and maximizes the overall effectiveness of the methodology.

Regarding the second stage, where the clustering algorithms were applied, the existence of mixed variables, with quantitative and qualitative values and the dependence of some variables, made it necessary to carry out a comprehensive analysis of the data set and some transformations of these features. The analysis of the identified typologies for their formal definition reflected a clear dominion of the qualitative variables in the formation of clusters, what leads to think if this would happen in any data set with quantitative and qualitative variables.

Finally, the applied clustering techniques produced much better results than expected, with an insignificant number of possible assignment errors. From the geostatistical analysis of the entire municipality, it was found that the data estimation through the clustering results was more updated than what could have been obtained from the available orthophotographs for 70% of the parcels in the studied municipality. Moreover, the proposed methodology allows to automatically update this information with each update of the Cadastre database. The Ward's method was clearly the most adequate for the purposes of the study, obtaining quality metrics that automatically discarded any possibility of improvement through another hierarchical clustering method. In addition, the identification of medoids would allow using them as representative case studies within their respective clusters and, therefore, to generalize results to the entire cluster to which they belong. This strategy is expected to be useful for analysing big sets of buildings at urban scale in order to obtain benchmarks for each typology that allow comparing and evaluating results from new case studies. Future developments would have to apply a complete analysis of the feasibility of energy retrofitting actions to a municipality according to the identified typologies.

The explained methodology presents three main limitations. First, the original data set from the GDC contains registers with empty or wrong data whose treatment cannot always be automated since the form of errors might be unpredictable in some cases. These can only be characterized through an initial data set inspection and the analysis of the existing values for each feature, and this must be a manual process. Second, the addition of other energy-related features such as façades orientation or shape factors was not feasible with the current data set, where the parcel is the smallest item geographically identified. In case buildings and their shape were identified in the future, it would be necessary to match the buildings identified by address in this study with those represented in the map, and a complex geometric analysis should be carried out to obtain orientations and dimensions. Finally, the dominion of qualitative variables in the results, such as the period of construction or the type of building allowed a generalization of features within each typology to make possible the analysis at the urban-scale. However, the stronger this generalization is, the lower the precision of the estimation becomes, so a balance between these two aspects should be pursued in forthcoming research.

CRediT authorship contribution statement

Alejandro Martínez-Rocamora: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Pilar Díaz-Cuevas:** Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Juan Camarillo-Naranjo:** Data curation, Formal analysis, Investigation, Software. **David Gálvez-Ruiz:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Patricia González-Vallejo:** Data curation, Investigation, Methodology, Validation, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Alejandro Martinez-Rocamora reports financial support was provided by General Secretariat of Housing of the Andalusian Ministry of Development, Articulation of Territory and Housing. If there are other authors, they declare that they have no known competing

financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This publication is part of the R&D project US.22–08 (ÁGORA: Análisis geográfico predictivo económico y ambiental de operaciones de mantenimiento y rehabilitación del parque residencial público de Andalucía), funded by the Andalusian Government through the General Secretariat of Housing of the Andalusian Ministry of Development, Articulation of Territory and Housing, and has also been possible thanks to the VI Own Research and Knowledge Transfer Plan from the University of Seville (VI-PPIT US).

References

[1] L.G. Swan, V.I. Ugursal, Modeling of end-use energy consumption in the residential sector: a review of modeling techniques, Renew. Sustain. Energy Rev. 13 (2009) 1819–1835, https://doi.org/10.1016/j.rser.2008.09.033.
[2] C. Rivero-Camacho, J.J. Martín-del-Río, M. Marrero-Meléndez, Evolution of the life cycle of residential buildings in Andalusia: economic and environmental evaluation of their direct and indirect impacts, Sustain. Cities Soc. 93 (2023) 104507, https://doi.org/10.1016/j.scs.2023.104507.
[3] P. González-Vallejo, R. Muntean, J. Solís-Guzmán, M. Marrero, Carbon footprint of dwelling construction in Romania and Spain. A comparative analysis with the OERCO2 tool, Sustain. Times 12 (2020), https://doi.org/10.3390/SU12176745.
[4] R.J. Muringathuparambil, J.K. Musango, A.C. Brent, P. Currie, Developing building typologies to examine energy efficiency in representative low cost buildings in Cape Town townships, Sustain. Cities Soc. 33 (2017) 1–17, https://doi.org/10.1016/j.scs.2017.05.011.
[5] T.A. Bailey, R. Dubes, Cluster validity profiles, Pattern Recogn. 15 (1982) 61–83, https://doi.org/10.1016/0031-3203(82)90002-4.
[6] M. van de Velden, A.I. D'Enza, M. Yamamoto, Special feature: dimension reduction and cluster analysis, Behaviormetrika 46 (2019) 239–241, https://doi.org/10.1007/s41237-019-00092-6.
[7] A. Nowak-Brzezińska, I. Gaibei, How the outliers influence the quality of clustering? Entropy 24 (2022) https://doi.org/10.3390/e24070917.
[8] H. Naganathan, W.O. Chong, X. Chen, Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches, Autom. ConStruct. 72 (2016) 187–194, https://doi.org/10.1016/j.autcon.2016.08.002.
[9] D. Bienvenido-Huertas, D. Marín-García, M.J. Carretero-Ayuso, C.E. Rodríguez-Jiménez, Climate classification for new and restored buildings in Andalusia: analysing the current regulation and a new approach based on k-means, J. Build. Eng. 43 (2021) 102829, https://doi.org/10.1016/j.jobe.2021.102829.
[10] A. Hollberg, T. Lützkendorf, G. Habert, Top-down or bottom-up? – How environmental benchmarks can support the design process, Build. Environ. 153 (2019) 148–157, https://doi.org/10.1016/j.buildenv.2019.02.026.
[11] A. Martínez-Rocamora, C. Rivera-Gómez, C. Galán-Marín, M. Marrero, Environmental benchmarking of building typologies through BIM-based combinatorial case studies, Autom. ConStruct. 132 (2021), https://doi.org/10.1016/j.autcon.2021.103980.
[12] R. Frischknecht, M. Balouktsi, T. Lützkendorf, A. Aumann, H. Birgisdottir, E.G. Ruse, A. Hollberg, M. Kuittinen, M. Lavagna, A. Lupišek, A. Passer, B. Peuportier, L. Ramseier, M. Röck, D. Trigaux, D. Vancso, Environmental benchmarks for buildings: needs, challenges and solutions—71st LCA forum, Swiss Federal Institute of Technology, Zürich, 18 June 2019, Int. J. Life Cycle Assess. 24 (2019) 2272–2280, https://doi.org/10.1007/s11367-019-01690-y.
[13] U. Ali, M.H. Shamsi, M. Bohacek, C. Hoare, K. Purcell, E. Mangina, J. O'Donnell, A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings, Appl. Energy 267 (2020) 114861, https://doi.org/10.1016/j.apenergy.2020.114861.
[14] E.G. Dascalaki, K.G. Droutsa, C.A. Balaras, S. Kontoyiannidis, Building typologies as a tool for assessing the energy performance of residential buildings - a case study for the Hellenic building stock, Energy Build. 43 (2011) 3400–3409, https://doi.org/10.1016/j.enbuild.2011.09.002.
[15] S. Ghanbari, M. Yeganeh, M. Reza bemanian, Architecture typology of rural plain houses based on formal features, case study: (Talesh, Iran), Front. Built Environ. 8 (2022) 1–15, https://doi.org/10.3389/fbuil.2022.856567.
[16] C. Jiménez-Pulido, A. Jiménez-Rivero, J. García-Navarro, Caracterización de fachadas: clasificación de las tipologías constructivas más habituales en España, Inf. La Construcción. 74 (2022) e471, https://doi.org/10.3989/ic.88694.
[17] É. Mata, A. Sasic Kalagasidis, F. Johnsson, Building-stock aggregation through archetype buildings: France, Germany, Spain and the UK, Build. Environ. 81 (2014) 270–282, https://doi.org/10.1016/j.buildenv.2014.06.013.
[18] L. Pistore, G. Pernigotto, F. Cappelletti, A. Gasparella, P. Romagnoni, A stepwise approach integrating feature selection, regression techniques and cluster analysis to identify primary retrofit interventions on large stocks of buildings, Sustain. Cities Soc. 47 (2019) 101438, https://doi.org/10.1016/j.scs.2019.101438.
[19] E. Rodrigues, D. Sousa-Rodrigues, M. Teixeira de Sampayo, A.R. Gaspar, Á. Gomes, C. Henggeler Antunes, Clustering of architectural floor plans: a comparison of shape representations, Autom. ConStruct. 80 (2017) 48–65, https://doi.org/10.1016/j.autcon.2017.03.017.
[20] X. Shan, Q. Deng, Z. Tang, Z. Wu, W. Wang, An integrated data mining-based approach to identify key building and urban features of different energy usage levels, Sustain. Cities Soc. 77 (2022) 103576, https://doi.org/10.1016/j.scs.2021.103576.
[21] Y. Wu, B. Mashhoodi, A. Patuano, S. Lenzholzer, L. Narvaez Zertuche, A. Acred, Heat-prone neighbourhood typologies of European cities with temperate climate, Sustain. Cities Soc. 87 (2022), https://doi.org/10.1016/j.scs.2022.104174.
[22] Efinovatic, CE3X v2.3 - Software for Energy Certification of Buildings, 2023. http://www.efinova.es/CE3X.
[23] TABULA Project Team, TABULA WebTool, 2017. https://webtool.building-typology.eu/#bm. June 30, 2023).
[24] T. Loga, B. Stein, N. Diefenbach, TABULA building typologies in 20 European countries—making energy-related features of residential building stocks comparable, Energy Build. 132 (2016) 4–12, https://doi.org/10.1016/j.enbuild.2016.06.094.
[25] S. Hrabovszky-Horváth, T. Pálvölgyi, T. Csoknyai, A. Talamon, Generalized residential building typology for urban climate change mitigation and adaptation strategies: the case of Hungary, Energy Build. 62 (2013) 475–485, https://doi.org/10.1016/j.enbuild.2013.03.011.
[26] J.H. Ward Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244, https://doi.org/10.2307/2282967.
[27] Ministry of Finance and Civil Service, Sede Electrónica del Catastro, 2023. https://www.sedecatastro.gob.es/. April 10, 2023).
[28] General Directorate of Cadastre, Descarga y tratamiento de información alfanumérica en formato CAT, Manual del usuario., 2023. https://www.catastro.minhap.es/ayuda/manual_descargas_cat.pdf.
[29] General Directorate of Cadastre, Especificación de formato Shapefile, 2023. https://www.catastro.minhap.es/ayuda/manual_descriptivo_shapefile.pdf.
[30] R.T. Mora García, P. Marti Ciriquian, Desagregación poblacional a partir de datos catastrales, in: Análisis Espac. Y Represent. Geográfica, Innovación Y Apl., 2015, pp. 305–314. Zaragoza, Spain.
[31] J.M. Santos Preciado, La cartografía catastral y su utilización en la desagregación de la población. Aplicación al análisis de la distribución espacial de la población en el municipio de Leganés (Madrid), Estud. Geográficos 76 (2015) 309–333, https://doi.org/10.3989/estgeogr.201511.
[32] M.R. Anderberg, Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks, Academic Press, 2014.
[33] L. Kaufman, P.J. Rousseeuw, Partitioning around medoids (program PAM), in: Find. Groups Data an Introd. To Clust. Anal., John Wiley & Sons, Inc., 1990, pp. 68–125, https://doi.org/10.1002/9780470316801.ch2.
[34] J.C. Gower, A general coefficient of similarity and some of its properties, Biometrics 27 (1971) 857–871, https://doi.org/10.1109/ultsym.1987.199076.

[35] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for dicovering clusters in large spatial databases with Noise, in: KDD-96 2nd Int. Conf. Knowl. Discov, . Data Min., 1996, pp. 226–231, https://doi.org/10.11901/1005.3093.2016.318.
[36] G.N. Lance, W.T. Williams, A general theory of classificatory sorting strategies: 1. Hierarchical systems, Comput. J. 9 (1967) 373–380, https://doi.org/10.1093/comjnl/9.4.373.
[37] P.J. Rousseeuw, Silhouettes, A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.
[38] Ministry of Housing, Normas MV, 1957.
[39] Government of Spain, NBE-CT79: Normas Básicas de Edificación - Condiciones térmicas de los edificios/Basic Building Regulations - Thermal conditions of buildings, 1979. Spain, https://www.boe.es/boe/dias/1979/10/22/pdfs/A24524-24550.pdf.
[40] CTE, Spain MH (Ministry of Housing). Código Técnico de la Edificación (Building Technical Code), 2006. Madrid. Spain.
[41] R. R Core Team, A Language and Environment for Statistical Computing, 2021. Vienna, Austria, https://www.r-project.org/.
[42] Posit Team, RStudio, Integrated Development Environment for R, Boston, MA, 2023. http://www.posit.co/.
[43] IVE, Catálogo de tipología edificatoria residencial. Ámbito, España/Residential Building Typologies Catalogue, Region, Spain, Valencia, 2016. https://episcope.eu/building-typology/country/es/.
[44] M.D. Noguero Hernández, M. del P. Díaz Cuevas, J. Ojeda Zújar, Georreferenciación y Análisis Espacial de Actuaciones de Rehabilitación Residencial en Alcalá de Guadaira (Sevilla), in: XVI Congr. Nac. Tecnol. La Inf. Geográfica, 2014, pp. 684–694. Alicante, Spain, https://dialnet.unirioja.es/servlet/catart?codigo=5426056.
[45] J.P. Pérez-Alcántara, M. del P. Díaz-Cuevas, J.I. Álvarez-Francoso, J. Ojeda-Zújar, Métodos de adscripción Y tratamiento espacial para La generación Y visualización de indicadores de vivienda (grid) a través de Catastro, in: XVII Congr. Nac. Tecnol. Inf. Geográfica, 2016, pp. 224–234. Málaga, Spain.
[46] J.P. Pérez-Alcántara, J. Ojeda-Zújar, M. del P. Díaz-Cuevas, J.I. Álvarez-Francoso, Integración de datos poblacionales y catastrales en estructuras GRID : primeros resultados para el espacio residencial en Andalucía, in: XXV Congr. La Asoc. Geógrafos Españoles, 2017, pp. 1619–1628. Madrid, Spain.
[47] J. Camarillo, I. Vallejo, A. Tabales, E. Pavón, Where is tourist housing actually located? New approaches and sources for detailed scale analysis, Eur. Plann. Stud. 30 (2021) 1–25, https://doi.org/10.1080/09654313.2021.2002825.
[48] J.M. Martín Jiménez, V.M. Rodríguez Espinosa, Processing of the cadastre information using postgresql-postgis. Application to the analysis of urban land uses in Alcalá de Henares, Spain, Estud. Geográficos 83 (2022), https://doi.org/10.3989/ESTGEOGR.2022106.106.