Application of Data Mining Techniques to rural vernacular buildings: a methodology for characterisation and awareness

Catarina P. Mouraz[a]*, Ricardo M.S.F. Almeida[b], Tiago Miguel Ferreira[c], J. Mendes Silva[d]

[a] *University of Coimbra, Department of Civil Engineering, Coimbra, Portugal;*

[b] *Polytechnic Institute of Viseu, School of Technology and Management, Department of Civil Engineering; CONSTRUCT-LFC, University of Porto, Faculty of Engineering (FEUP)*

[c] *University of the West of England (UWE Bristol), College of Arts, Technology and Environment - School of Engineering, Bristol, UK*

[d] *University of Coimbra, ADAI, Department of Civil Engineering*

* University of Coimbra, Department of Civil Engineering, Rua Luís Reis Santos, Pólo II, 3030-788, Coimbra, Portugal; catarinamouraz@hotmail.com

Application of Data Mining Techniques to rural vernacular buildings: a methodology for characterisation

Data-mining techniques (DMTs) have been widely used in the context of existing buildings. However, studies comparing its application to different objects and using rural territories with vernacular constructions as case studies remain highly unexplored. This paper discusses a methodology applied towards characterising two rural settlements in Portugal and Spain using DMTs and GIS-based maps to support the interpretation of results. Conclusions such as the existence of outliers or the optimal number of clusters are drawn, as well as the nature of clusters and their pertinence in identifying buildings closer to the vernacular matrix. Adapting the numerical database is suggested as future research, mainly reducing the number of variables to obtain more accurate insights into the identification of vernacular buildings. Investigations such as the one presented in this paper promote awareness concerning the conservation and improvement of rural settlements, ultimately stimulating action towards territorial cohesion and sustainable development through future rehabilitation actions in vernacular built heritage.

Keywords: vernacular buildings; building rehabilitation; rural settlements; data mining techniques; cluster analysis; GIS.

1. Introduction

The application of data mining techniques (DMTs) in the field of civil and building engineering is not new. Even though current research is mostly focused on energy-performance concerns, where studies tackling energy consumption, usage, or savings have proliferated in the past years (Zhang *et al.*, 2018; Zhao *et al.*, 2020; He *et al.*, 2022), these techniques have a wide range of application due to their role in identifying previously unknown relationships (Shan *et al.*, 2022) and extract unobvious information from large datasets.

Cluster analysis remains one of the most common DMTs applied to existing buildings, providing information about how buildings can be grouped based on commonalities (Tardioli *et al.,* 2010), enabling researchers to deal with large datasets composed by heterogeneous and scattered information (Pistore *et al*., 2019). By classifying cases into groups, cluster analyses allow uncovering correspondences within data to create meaningful groups of objects (Yim *et al.*, 2015). These approaches can be applied to various scopes of interest: conservation state analyses (Rosti *et al.*, 2022), definition of building typologies and archetypes (Arambula Lara *et al.*, 2014), or development of intervention strategies and policies (Mouraz *et al.*, 2022a), to cite a few. The application of these techniques to large groups of buildings, especially considering different territorial scales, may allow obtaining new, useful, and otherwise unperceived insight regarding groups of elements with heterogeneous features. However, little research has been developed regarding the discussion of the effectiveness of applying DMTs in different-sized groups of buildings since most studies focus on cluster distribution within a set of buildings and not on the comparison of results between different sets of buildings with various dimensions.

Furthermore, the application of DMTs in the building sector remains almost exclusively focused on urban contexts (Afaifia *et al.*, 2021; Rosti *et al.*, 2022; Shan *et al.* 2022). The fact that rural residents make up 44% of the global population highlights the disproportionate gap existing between these two realities (The World Bank, 2023), and research regarding buildings located in these territories remains considerably less explored and documented. Moreover, the current global phenomena of urbanisation and the consequent desertification of rural settlements directly counteract territorial cohesion and, ultimately, the global objective of sustainable development (Alexiadis, 2017). Even though more studies are arising regarding the retrofit of buildings in rural

territories, especially focusing on energy-performance concerns (Jiang *et al.*, 2022; Li *et al.*, 2022), there is still a compelling lack of research surrounding their characterisation as housing territories and the variety of buildings they entail remains highly undocumented.

Built vernacular heritage in rural areas is a particularly relevant part of these territories since it constitutes a physical testimony of a community's wisdom regarding the way of inhabiting a certain place (Nguyen *et al.*., 2019). The comprehensive approach embodied by these buildings as a response to an economic, environmental, and social context justifies the pertinence of considering them an example of sustainability (Mouraz *et al.*, 2023). Despite the growing interest in the conservation and relevance of certain features, especially in southern European countries (Elert *et al.*, 2021; Parracha *et al.*, 2021), research is still scarce regarding the application of methodologies that consider vernacular constructions as valid case studies. Promoting the characterisation of rural environments that include vernacular-built heritage tackles a profound gap in research and raises awareness of the conservation of these settlements as a catalyst towards global sustainable development (Yin *et al.*, 2019; Mouraz *et al.*, 2023). Furthermore, using them as case studies in the application of methodologies and tools widely used in the characterisation of urban contexts, such as DMTs, remains a gap of research which is highly unexplored. The use of these techniques in favour of a deeper characterisation of rural territories with vernacular heritage poses, therefore, a fundamental added value in research due to their efficiency in obtaining valuable knowledge from large datasets (Zhao *et al.*, 2020). This information is useful to support more informed decision-making processes regarding future rehabilitation or conservation actions, as well as defining territorial-scale policies (Mouraz *et al.*, 2022a).

Considering the context mentioned above, it is clear there is a general lack of research surrounding the characterisation of rural settlements with vernacular construction, especially in the field of DMTs. Thus, considering two rural settlements with vernacular construction in Portugal and Spain, the research presented in this paper aims to achieve the following objectives:

- Contribute to the construction and architectural characterisation of rural settlements, raising awareness of their specific features, dimensions, and contexts;
- Validate the existence of buildings with characteristics closer to the vernacular matrix through cluster analysis;
- Discuss the pertinence, usefulness, and constraints of applying DMTs to different objects in the context of rural settlements with vernacular construction, reviewing results obtained for different objects.

The expected outcomes are the characterisation of buildings in these rural settlements, mainly identifying groups with similar characteristics, as well as outliers. The developed research represents an innovative approach due to three main reasons: the consideration of rural settlements as valid case studies, especially considering widely used methodologies in urban contexts such as DMTs; the discussion of the application of these techniques in different sized samples, contrary to common practice which usually is focused on the analysis of results within the same set of buildings; the combined use of DMTs that are usually individually applied in the field of outlier detection.

Our findings deepen the knowledge of rural territories and constructions, bearing two main positive consequences: supporting more informed rehabilitation actions and

territorial-scale policies; acting as a tool of awareness towards the relevance of vernacular buildings as proven models of a sustainable response in the built environment (Mouraz *et al*., 2023).

2. Methodology

The methodology adopted in this study involves three steps illustrated in Figure 1: data collection for case study characterisation, the application of DMTs to data sets, and GIS visualisation of the cluster analysis results. A detailed description of the methodology steps is presented in subsections below.

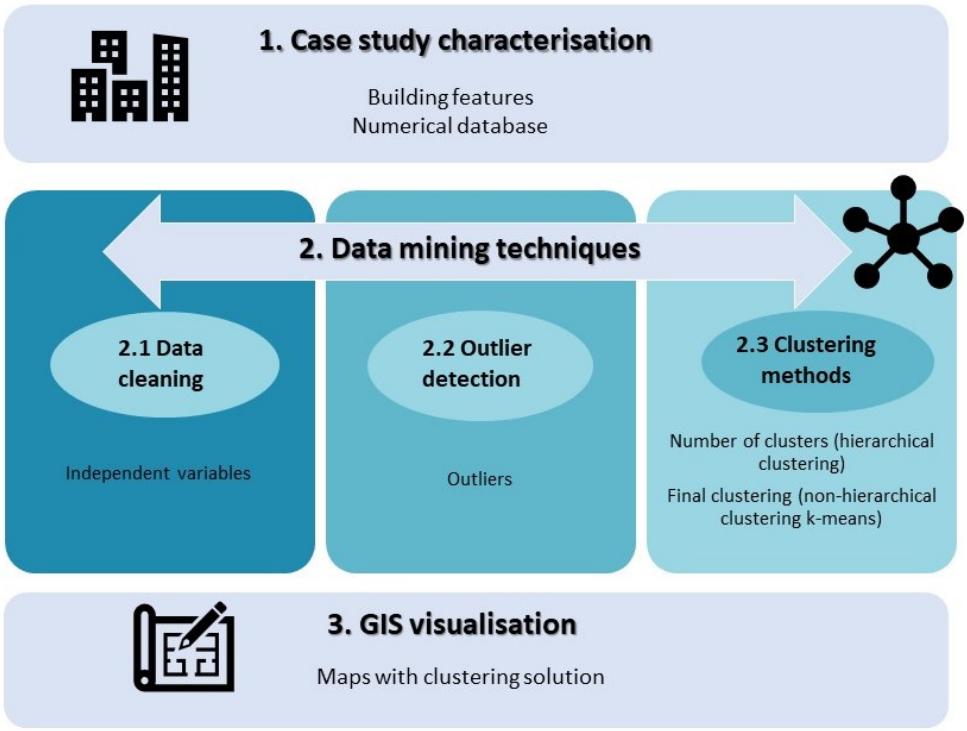


Figure 1. Research framework adopted in this study: tasks and corresponding outputs.

2.1 Case study characterisation

Two case studies were investigated in this research: Chiqueiro, a mountain settlement located in Lousã mountains (Serra da Lousã), central Portugal (case study 1), and Rello,

a rural village in Soria Province, in Spain (case study 2). The first step of the methodology was the characterisation of these case studies, achieved through three complementary steps: survey framework preparation; fieldwork; and off-site work for data processing, analysis, and development of a database (Vicente *et al.*, 2015).

Survey preparation consisted of the organisation of procedures and tasks related to the on-site survey. Considering the goals of this research, identified in Section 1, the scope of the analysis was defined, encompassing mainly the construction and architectural elements necessary to observe on-site.

Fieldwork consisted of an on-site survey in which the buildings were thoroughly inspected and photographed from the outside, especially the elements defined as essential in the previous step, such as masonry walls, roofs, or window framings, allowing to collect necessary information in a coherent and organised manner. Also, video footage was recorded in both case studies, either resorting to hand cameras or unmanned aerial vehicles, which were used to latter obtain information regarding hardly accessible elements of buildings, such as roofs.

Finally, off-site work was conducted to create a structured database that included the significant volume of information collected on the *in situ* survey. Notes and observations registered on-site, collected photographs, and frames from the video footage allowed for the thorough characterisation of buildings and their construction elements.

Considering the buildings' construction and architectural features, the collected information was organised according to three groups of features:

- Group A: Geometry (ex.: number of floors; percentage of openings area in the main façade; roof shape).
- Group B: Functionality and habitability (ex.: use; solar orientation; level of habitability and state of conservation).

- Group C: Technology and materials (ex.: window framing material; type of façade coating; lintels material).

To enable the application of DMTs, which are mainly focused on numerical analysis, the developed databases were populated using a binary system considering all possible features as variables, in which "0" means the building does not present a particular feature, and "1" is for the contrary, allowing for a direct classification of the presence or absence of characteristics.

2.2 DMTs

Previous works

DMTs have been extensively applied in the building construction field for different purposes. Clustering methods is one of the most widely applied techniques, especially for the analysis of large groups of buildings and substantial amounts of data. Table 1 presents a set of works analysed for the purpose of this research, summarising goals, objects of study, and DMTs applied, such as data cleaning, pre-clustering and clustering methods, validation steps, and outlier detection.

Table 1. Research works reporting the application of DMTs to groups of buildings.

Reference	Goal of the research	Number analysed buildings	Data cleaning (Y/N)	Pre-clustering method	Clustering method	Validation step	Outlier detection (Y/N)
Arambula Lara et al., 2014	Characterize the features and performance of the buildings, and to assess the possibility to select a sample of representative schools to be further monitored	49	Y	-	k-means	Linear regressive models	N
Schaefer et al., 2016	Obtain reference buildings for the low-income housing stock	120	N	Hierarchical clustering (find optimal k and exclude outliers)	k-means	-	Y
Li *et al.*, 2018	Find building archetypes; residential building stock model for energy calculations	575	Y	nbClust (find optimal k)	K-means; K-medoids	-	N

Patteeuw *et al.* 2018	Present the application of a method called cluster-centre aggregation (CCA) in building stock simulation and evaluate its performance	2098	N	-	Hierarchical clustering (dendrogram analysis)	-	N
Pistore *et al.*, 2019	Describe and classify a set of existing buildings from an energy perspective in order to highlight possible solutions to improve their energy performance and to establish a priority list of intervention	41	Y	Hierarchical cluster (find optimal k and exclude outliers)	k-medoids	-	Y
Mouraz et al., 2022a	Propose a methodology which extracts additional information from a database on the state of conservation of building stock, combining the use of cluster analysis and GIS maps.	495	N	-	Hierarchical clustering	Silhouette	N
Cheng *et al.*, 2023	Evaluate spatial sequences represented by building facades in regional traditional villages	325	Y	-	Hierarchical clustering (dendrogram analysis)	-	N

From the analysis of Table 1, one can observe that published research on this topic has focused on studying large sets of buildings, yet with varied dimensions – from tens (Pistore *et al.,* 2019) to thousands of buildings (Patteeuw *et al.*, 2018). Data cleaning is used in several studies, highlighting its disseminated importance towards more reliable results.

Combining two clustering algorithms is also common, especially hierarchical algorithms, used to find the optimal number of clusters k, followed by non-hierarchical algorithms, such as k-medoids (in Pistore *et al.,* 2019) or k-means (in Schaefer *et al.*, 2016), to obtain final clustering solutions. Both hierarchical and non-hierarchical methods are used with the same objective of enhancing the homogeneity of a certain group of objects, as well as the heterogeneity between these groups of objects (Yim *et al*., 2015). However, while hierarchical clustering resorts to progressively gathering similar elements in sequential iterations, non-hierarchical algorithms assign objects considering a set of centres that are initially established (Gulagiz *et al.*, 2017).

Finally, validation steps and the identification of outliers seem to be more rarely conducted. Studies presented in Table 1 considered either the analysis of dendrograms

(Pistore *et al.*, 2019) or the calculation of the Mahalanobis distance (Schaefer *et al.*, 2016) to perform this task, with the second approach being used in a larger set of buildings.

It is also clear that most studies focusing on the application of DMTs to large groups of buildings are conducted in urban environments (Li *et al.*, 2018; Schaefer *et al.*, 2016; Mouraz *et al.*, 2022), opposing to rural territories with vernacular or traditional constructions (Cheng *et al.*, 2023).

Based on this information, the methodology employed in this study follows a systematic approach, as detailed and justified in the subsequent sections. The methodology encompasses several key steps, starting with data cleaning to ensure data quality and integrity. Subsequently, outlier detection techniques are applied to identify and handle any anomalous data points. Finally, cluster analysis is conducted to group and characterise the rural settlements based on commonalities found on the buildings that compose them.

Data cleaning

Data cleaning is often the first step in numerical analyses, due to its role in reducing error and improving the quality of data (Javatpoint, 2023), ultimately contributing to more accurate, coherent, and valid results. By highlighting important information, data cleaning allows for extracting hidden relationships from large data sets and getting more accurate inputs for further research steps, namely the application of clustering algorithms (Tardioli *et al.*, 2018).

Considering both the goal of developing a cluster analysis and putting together a database of variables obtained during the building characterisation stage, data cleaning was a fundamental procedure. Thus, a variable correlation analysis was conducted since highly correlated variables can negatively influence clustering results (Li *et al.*, 2018). When variables are perfectly correlated, they represent the same concept, influencing the

final solution, which will necessarily consider the same variable twice (Sambandam, 2003). Reducing the original data set to a small group of unequivocally independent variables contributes, thus, to more accurate results.

To do so, the similarity between variables is calculated through an absolute value of resemblance between them based on a numerical measure, such as the Euclidean distance or Pearson's correlation (Iglesias *et al.*, 2013).

Given the wide use of the latter measure, especially in numerical variables (Nettleton, 2014), Pearson's correlation was adopted in this work and calculated through SPSS software. This value ranges from -1 to 1, where interval ends indicate a strong negative correlation (-1) or strong positive correlation (1), whilst values close to 0 are associated with weak correlations (Li *et al.*, 2018).

Obtaining a correlation matrix through SPSS software allowed for the identification of strongly and weakly correlated variables considering the obtained values for "Sig. (2-tailed)" parameter. In SPSS, this parameter can be used to identify significant correlations between two variables when the value obtained is less than 0,05 (Obilor *et al.*, 2018). Only strong correlations were considered and analysed, leading to the elimination of redundant variables from the original database and to the creation of a smaller data set which was used in subsequent steps of the methodology.

Outlier detection

Outlier detection in the multivariate analysis allows identifying objects that differ from the remaining elements of a sample (Mayrhofer *et al.*, 2023), leading to the detection of unusual behaviour or characteristics. Considering the research goal of validating the existence of vernacular characteristics in buildings in rural settlements, outlier detection is a fundamental step in the analysis. It enables the immediate recognition of buildings

that do not align with the primary focus of the study, both in terms of construction and architecture.

Multivariate analyses require multivariate techniques in the identification of outliers (Schaefer *et al.*, 2016) and, for that end, the calculation of the Mahalanobis distance and the associated probability is a method commonly used in various fields of knowledge (Schaefer *et al.*, 2016; Yan *et al.*, 2018; Mayrhofer *et al.*, 2023). This parameter measures the distance of each object (building) to the centroid of the distribution that represents the sample (multivariate average), considering its covariance (multivariate variance) (Schaefer *et al.*, 2016), as shown in Eq. (1). The probability associated with Mahalanobis distance is used as a decisive parameter in numerical analyses, and objects with values under 0.001 are considered outliers (Schaefer *et al.*, 2016). For an observation vector $x = (x_1, \ldots, x_p)'$ from a population with expectation vector $\mu = (\mu_1, \ldots, \mu_p)'$ and covariance matrix C, the squared Mahalanobis distance of x to μ with respect to C is given as:

$$MD^2_{\mu,C}(x) = (x - \mu)' \cdot C^{-1} \cdot (x - \mu) \quad (1)$$

There are also other methods of detecting outliers, such as cluster-based approaches. These comply with more intuitive methodologies in which data that results from the clustering algorithm are analysed (such as dendrograms), and objects can be considered outliers either if they appear too distant from data clusters or if they integrate a smaller cluster of elements (Ray *et al.*, 2022).

In this paper, the identification of outliers is expected to allow the identification of buildings that are dissonant from the vernacular matrix (typically requiring a specific approach in terms of conservation strategies) and, thus, reduce the sample to the true object of the study. To that end, an innovative methodology based on the combination of

the two methods is proposed, where the calculation of the Mahalanobis distance probability is the first step of the outlier detection process, followed by the analysis of dendrograms obtained through hierarchical clustering. Given the findings from previous research on the use of these techniques in studies with samples of varying sizes and the distinctive characteristics of both study cases, particularly in terms of the number of buildings involved, the combined use of these techniques is proposed to analyse their effectiveness in outlier detection in both study cases.

Cluster analysis

Cluster analysis was then carried out for both case studies, excluding outliers and using databases composed of independent variables. Like other research works (Schaefer *et al.*, 2016; Li *et al.*, 2018; Pistore *et al.*, 2019), the cluster analysis approach followed a two-step process in which hierarchical and non-hierarchical clustering algorithms were combined to reach an optimised solution.

Hierarchical algorithms are suggested to be used as the primary step because they allow obtaining the indication of the optimal number of clusters (K) contained in the dataset. In this research, Squared Euclidean Distance was selected in SPSS as the parameter to measure the similarity between elements, and between-groups linkage was selected as the agglomerative hierarchical clustering method. To identify K, the visual analysis of the dendrogram (combined with the numerical analysis of the agglomeration schedule for validation due to its role in identifying at which point two clusters which are being combined are increasing in heterogeneity through the increase in coefficient values) was used as "the stopping rule" (Yim *et al.*, 2015).

Then, the non-hierarchical algorithm was applied considering the K number of clusters provided by the previous step to obtain the final clustering solution. K-means was the chosen algorithm, a partitional clustering technique (non-hierarchical) that

assigns objects to clusters by minimising distances between these objects and the centroids of the clusters (Li *et al.*, 2018). The formation of clusters is an iterative process that stops when the assignment of objects to each centroid remains unchanged. These techniques require the initial input of a user-defined K, which highlights the importance of first applying the hierarchical algorithm described in the previous step.

Following the examples of previous researchers (Li *et al.*, 2018; Pistore *et al.*, 2019; Schaefer *et al.*, 2016 apud Hair *et al.*, 2009; Bussab *et al.*, 1990), the combination of a hierarchical and non-hierarchical techniques is used in the following manner: 1. The hierarchical algorithm is only used to obtain the optimal number of clusters k, since the agglomerative process is irreversible, which means that objects cannot be iteratively reassigned (Yim *et al.*, 2015); 2. The non-hierarchical algorithm is then used to obtain the final clustering solution, since these methods require the definition of the number of clusters in the optimal solution and then rely on an iterative process.

Finally, for each case study, the formed clusters and their features were analysed. Hypotheses are drawn regarding the nature of each cluster, and the identification of possible typologies of buildings is carried out.

2.3 GIS visualisation

The final step of the followed methodology consisted of the graphical visualisation of results obtained in the cluster analysis. As in previous works by the authors (Mouraz *et al.*, 2022a), a GIS-based software (QGIS) was used to map the clustering solution obtained in each case. The visualisation and analysis of these maps may provide helpful information regarding the spatial distribution of buildings belonging to the same cluster, shedding some light on the interpretation of clustering results by identifying possible patterns in buildings with similar characteristics.

3. Case study 1. Chiqueiro, Lousã mountains, Portugal

3.1 Characterisation

The first case study is Chiqueiro, a rural village located in Lousã mountains (Serra da Lousã), in the region of Coimbra, central Portugal (Figure 2). Chiqueiro integrates a group of villages scattered over the northwest sector of Lousã mountain, known as "Schist villages", due to the predominance of schist as construction material (Mouraz *et al.*, 2022b). Human occupation in the village dates to the 16th century, but migration movements since the 19th century have led to a profound decrease in permanent population levels. Nowadays, the permanent population is extremely scarce, and few houses are temporarily occupied in summer (Mouraz *et al.*, 2022b).

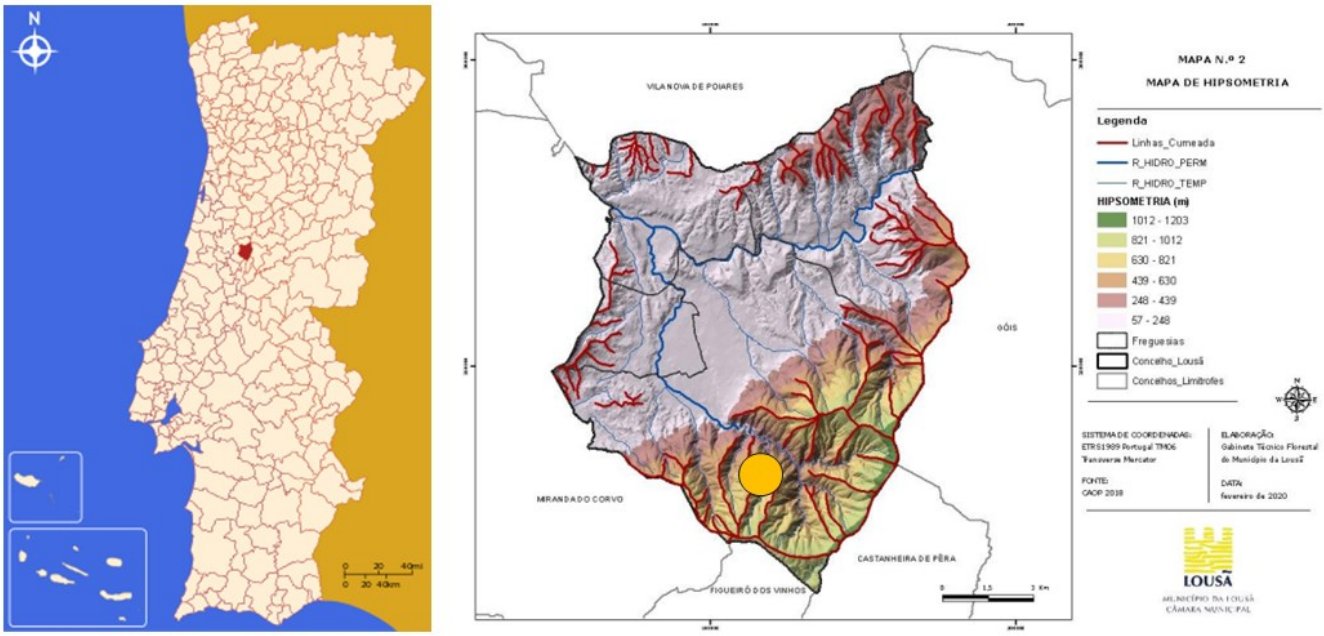


Figure 2: (1) Location of Lousã municipality in the Portuguese territory (Wikipedia, 2023a) and (2) location of Chiqueiro (in yellow) on the Lousã's hypsometric map (adapted from Município da Lousã, 2020)

Chiqueiro is located on the mountain's western slope and is composed of around 30 buildings organised in terraces following the hill's topography. The past importance of pastoralism justifies the predominance of two-floored buildings used for both housing and animal sheds (Carvalho, 2009). Figure 3 shows illustrative examples of Chiqueiro's vernacular buildings.

The exterior survey allowed the collection of information on the primary construction and architectural characteristics of buildings. With regards to their geometry (group A of features), it was observed that most buildings have two floors and two-sloped gable-shaped roofs. Openings are usually middle-sized, occupying 10 to 25% of the main façade area, and floor plans are typically rectangular.

Aerial and ground views of the village

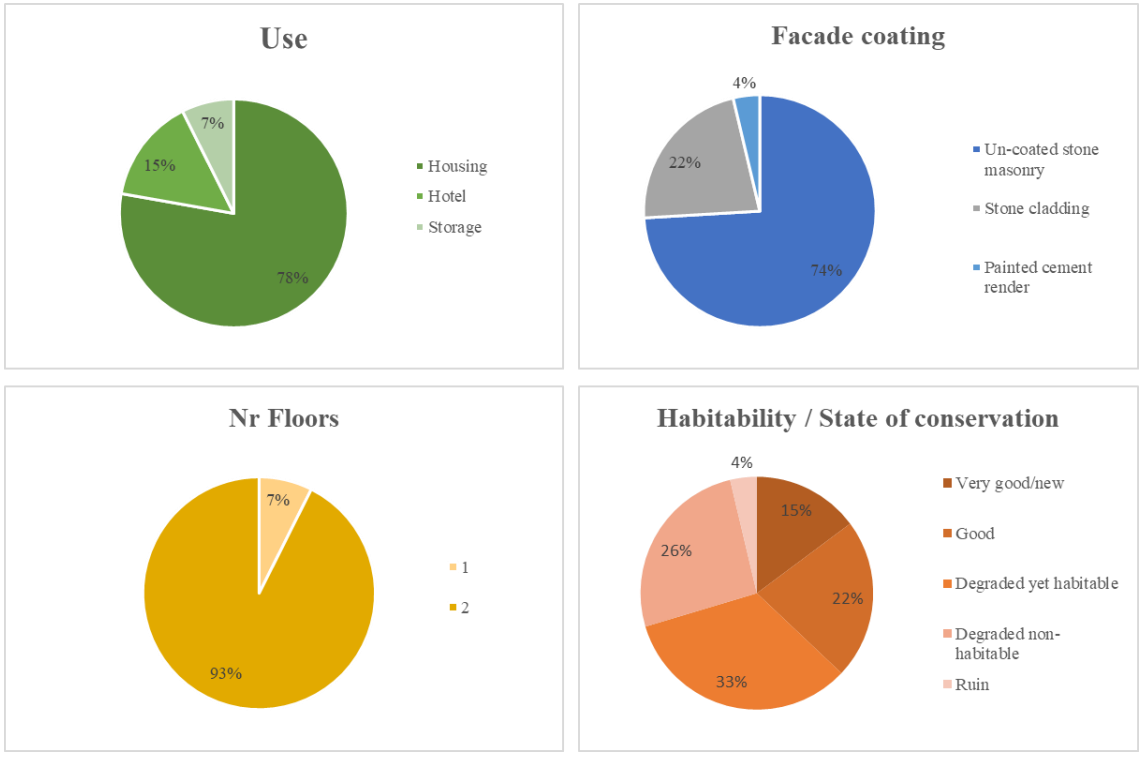
Schist masonry wall and clay tile coated roof

Figure 3. Illustrative examples of Chiqueiro's vernacular buildings (case study 1).

Considering functionality and habitability characteristics (group B of features, as per explained in Section 2), most buildings are used for housing, except for the local

church and a few buildings that are used for tourism accommodation or storage. Four buildings appear to have undergone renovation, despite remaining unfinished, and a considerable number of degraded buildings were identified.

Regarding technology and materials (group C of features), schist is the predominant construction material in masonry walls, which are mainly composed of



rubble-shaped stones with an uncoursed arrangement, following the masonry apparatus classification proposed by Szabó *et al.*, 2023. Dissonant coatings such as stone cladding or cement-based renders were also found, yet in a few cases and mainly in buildings that have undergone recent alterations. When observable, window framings were made of wood and roof coatings consisted mainly of clay tiles, especially barrel type, except for altered buildings which present Lusa-type clay tiles. In settlements like Chiqueiro, buildings with vernacular characteristics present construction elements built with local materials, such as schist and wood, available at the time of construction. Figure 4 compiles a set of statistical information characterising case study 1.

Figure 4. Statistical information characterising buildings in case study 1: use, number of floors, type of façade coating, and conservation state

Information collected on the survey for case study 1 was compiled and resulted in a numerical database composed of 61 variables and 29 buildings, where each variable corresponds to a feature observed regarding the set of concerns previously described.

3.2 Application of DMTs

As explained earlier, data preparation and cleaning were the first DMT applied to the database of information. The analysis of the correlation between variables identified more than 100 strong correlations between the 61 variables composing the database, which allowed for the creation of a simplified database composed of 26 independent variables (Table 2). Also, two buildings were excluded from the database (one for missing data and another currently used for religious purposes), resulting in 27 buildings.

Table 2- List of 26 independent variables obtained for case study 1

Group of features	Variable code	Description
A- Geometry	Var01	Number of floors: 1
	Var03	% of openings in main facade < 10%
	Var05	% of openings in main facade > 25%
	Var06	Roof shape: gable
	Var10	Number of roof slopes: 3
	Var11	Opening alignment: vertically aligned with the wall
	Var13	Opening alignment: non-applicable
B- Functionality and habitability	Var14	Floor-plan shape: rectangular
	Var17	Floor-plan shape: other
	Var20	Use: storage
	Var22	Occupancy: rarely occupied
	Var25	Solar orientation of main facade: north
	Var27	Solar orientation of main facade: east
	Var29	Access to upper floors: exterior stairs on main facade
	Var30	Access to upper floors: exterior stairs on secondary facade
	Var31	Access to upper floors: through the inside
	Var33	Access to upper floors: directly through patio

	Var35	Main entrance: directly through ground floor
	Var37	Main entrance: through private patio
	Var39	Private patio: yes
	Var42	Habitability/Conservation state: good
	Var44	Habitability/Conservation state: degraded non-habitable
C-Technology and materials	Var46	Window framing material: wood
	Var50	Facade coating material: painted cement render
	Var58	Lintels material: wood
	Var59	Lintels material: stone

Then, the calculation of Mahalanobis distance and respective probability did not highlight the existence of outliers since there were no probability values under 0.001. However, reporting to the cluster dendrogram obtained from the hierarchical algorithm as shown inFigure 5, it was possible to identify a group of 8 buildings, highlighted in this figure, that are included in the latter stages of the hierarchical process, showing a certain degree of heterogeneity regarding the remaining sample.

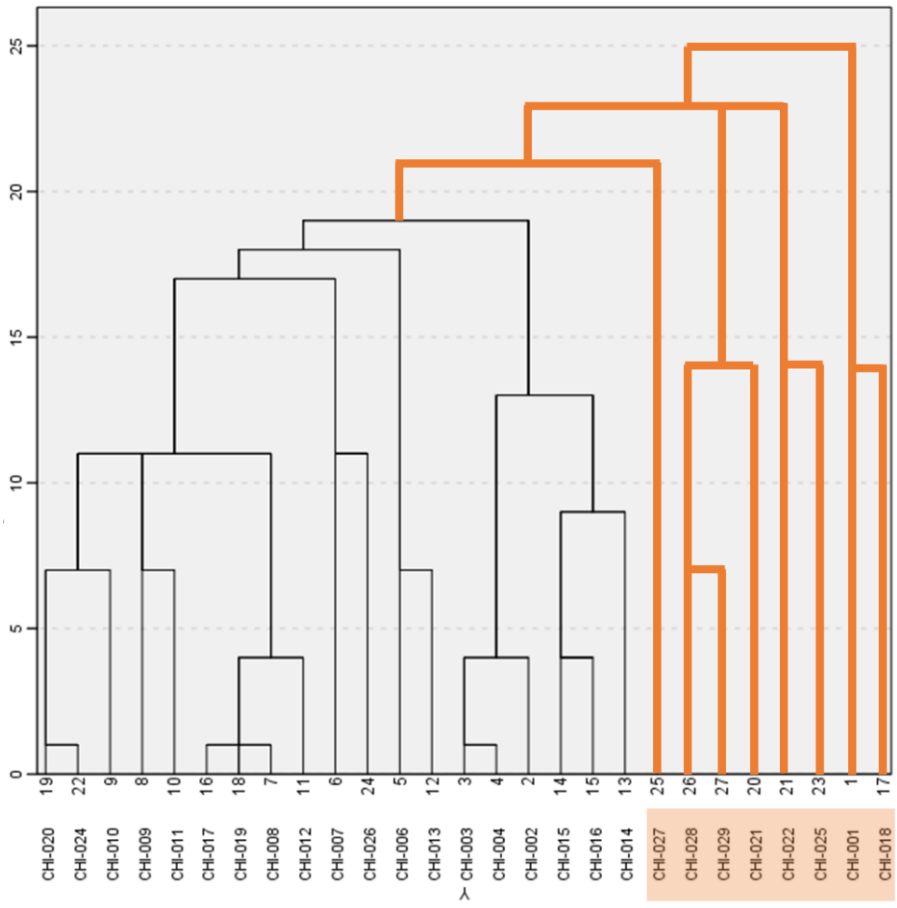


Figure 5. Cluster dendrogram obtained through hierarchical clustering algorithm for case study 1 and identification of outliers.

These buildings were considered case by case and analysed separately, leading to the decision to consider all outliers, either due to insufficient collected information or high levels of discrepancy regarding architectonic or construction characteristics.

The dendrogram analysis considering the remaining 19 buildings pointed the existence of 4 clusters (K). The last step of K-means clustering, performed considering K=4 and the remaining sample, allowed for the obtention of the final clustering solution. Table 3 combines the information obtained from K-means (namely the number of buildings in each cluster) with the analysis of the database characteristics for the buildings that belong to each cluster, allowing for their general description.

Table 3. General description of the final clustering solution for case study 1

Cluster	Number of buildings	General description and characteristics
1	9	**Uncoated stone masonry buildings, currently vacant and degraded.** Characteristics such as: 2-sloped roofs, rectangular-shaped floor plan, degraded or ruin state of conservation, wood window framings, clay tile (barrel type) as roof coating, wood lintels.
2	4	**Uncoated stone masonry housing buildings, rarely occupied and habitable (summer houses, group 1).** Characteristics such as: 2-sloped roofs, rectangular-shaped floor plan, main entrance in the upper floor, wood window framings, clay tile (barrel type) as roof coating, wood lintels.
3	4	**Buildings under construction.** Characteristics such as: 2-sloped roofs, North-orientated, without window framings (not yet placed), stone cladding as façade coating, clay tile (Lusa-type) as roof coating, wood lintels.
4	2	**Uncoated stone masonry housing buildings, rarely occupied with L-shaped floor plans (summer houses, group 2).** Characteristics such as: medium-sized openings in the main facade, 2-sloped roofs, North-orientated, good state of conservation, clay tiles (barrel type), wood lintels.

The final clustering solution and the differences between clusters appear to be coherent. Cluster 1 corresponds to uncoated stone masonry buildings, which are currently degraded or in ruins. Even though this group of buildings is composed of almost half of the total sample (9 out of 19), it appears to correspond to buildings with features closer to the vernacular matrix. Characteristics such as uncoated masonry wall apparatus, wood window framings, or clay barrel types as roof coating contribute to this conclusion, as well as the current conservation state of the buildings.

Clusters 2 and 4 include rarely occupied housing buildings, with the main difference being the floor-plan shape (rectangular-shaped in Cluster 2 and L-shaped in Cluster 4). These buildings may include summer houses, which are typical of these mountain settlements, and characterise their current occupation. However, the maintenance of certain construction features, such as uncoated stone masonry, clay barrel tiles, or wood lintels, may indicate the preservation of characteristics closer to traditional construction.

Moreover, Cluster 3 appears to differ significantly from the remaining three clusters. It is composed of four buildings currently under construction, in which characteristics such as stone cladding coating or Lusa-type tiles are examples of recent construction materials used, indicating profound changes in construction features.

3.3 GIS visualisation

Figure 6 represents the geographical distribution of the final clustering solution for case study 1 (K=4).

Figure 6. Mapping of the final clustering solution obtained for case study 1.

Specific characteristics of the described clusters are easily visualised on the map, such as L-shaped floor plans in Cluster 4 buildings or north-orientated facades in Cluster 3.

Buildings considered outliers seem to concentrate on the left side of the village, along the main street. However, no clear patterns can be identified from the obtained spatial distribution of buildings, nor can further conclusions be drawn from the nature of the clusters that compose the sample.

4. Case study 2. Rello, Soria province, Spain

4.1 Characterisation

The second case study is Rello, a medieval village located on top of a limestone cliff in Soria province, in north-western Spain (Figure 7), which is composed of a village and castle surrounded by a medieval wall dating to the 12th century.

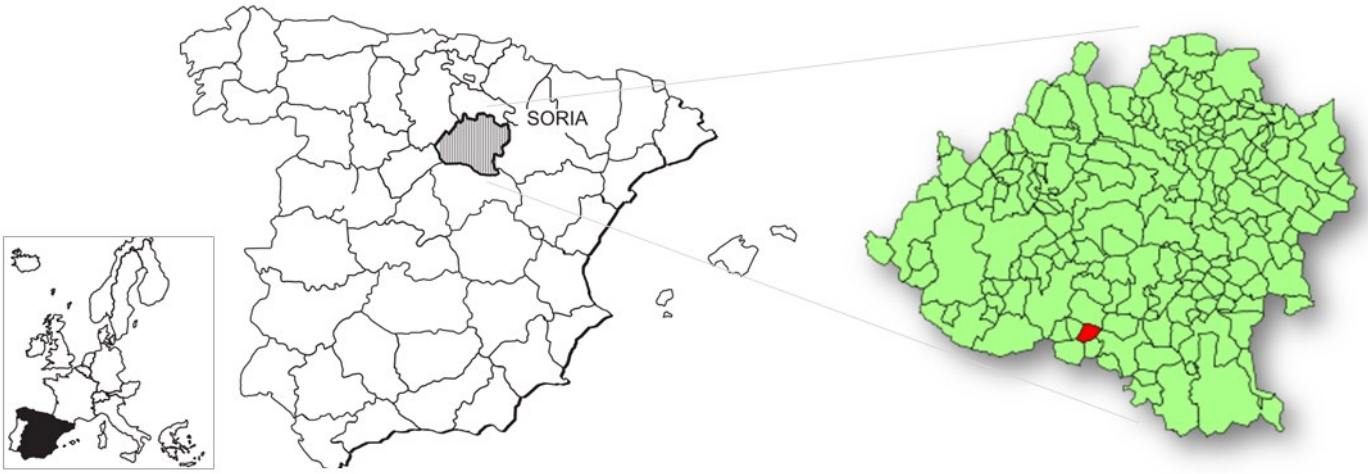


Figure 7: (1) Location of Soria province in Spain (adapted from Paniagua, 2008) and (2) location of Rello in Soria province (Wikipedia, 2023b)

Soria is currently one of the most scarcely populated territories of the European Union (European Commission, n.d.) due to an ongoing depopulation phenomenon since the 20th century. As for other villages in Soria and rural settlements in general, Rello has faced a depopulation challenge due to its territorial isolation and the inhabitants' pursuit of better living conditions throughout the centuries. The village develops at the same topographic level around three parallel streets and is composed of approximately 170 buildings, typically built in limestone (Figure 8).

Figure 8. Photographs of case study 2: Rello, Spain.

For this work, 128 buildings were surveyed from the outside (Figure 9). The developed survey allowed for observing that Rello's buildings are mostly one- or two-floor structures with gable or shed-shaped roofs (group A of features). Floorplans are mostly rectangular-shaped, with some cases presenting L or square shapes.

Considering group B of features, related with functionality and habitability concerns, buildings are mainly used for housing or garages. A significant number of one-floor buildings are currently vacant and presumably used for animal shedding, some of which are being used as storage spaces. Most buildings in Rello appear to be vacant, with a few rarely occupied. Information regarding conservation and habitability is coherent with these results, since most buildings were found to be in a degraded and non-habitable state of conservation.

Figure 9. Schematic view of the urban organisation of buildings in Rello (with analysed buildings in yellow)

Regarding technology and materials (group C of features), most façade walls are composed of uncoated stone masonry, with an uncoursed or brought-to-courses arrangement and rubble dressing. The most common roof coating material is clay tiles (barrel type), and there is a wide variety of lintel materials, with the predominance of limestone and wood elements. Figure 10 compiles a set of statistical information characterising case study 2.

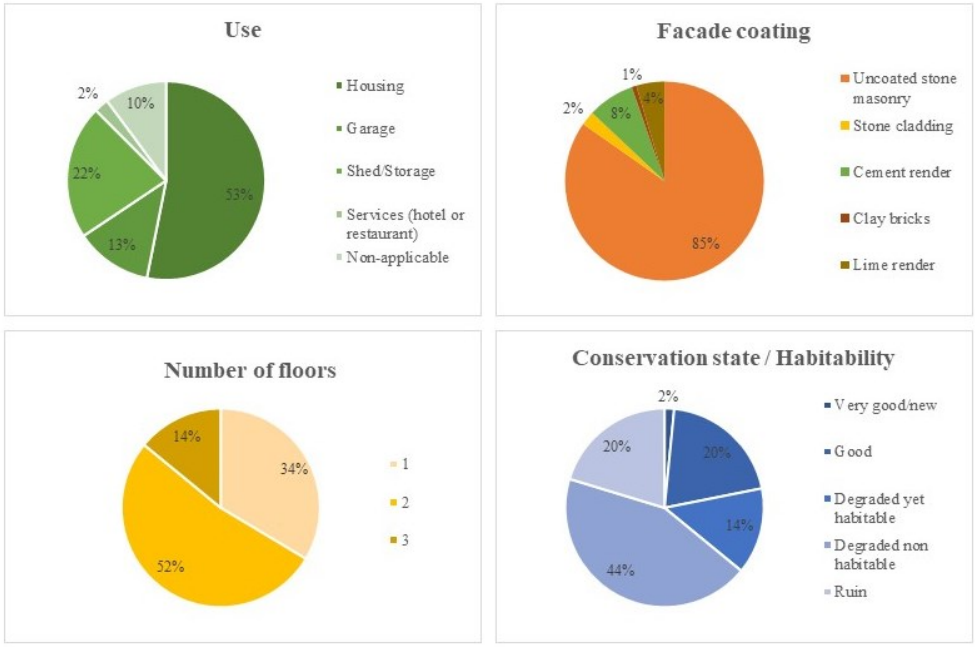


Figure 10. Statistical information characterising buildings in case study 2: use, number of floors, type of façade coating, and conservation state

Considering the same binary logic presented for case study 1, information collected on the survey for case study 2 was compiled on a database of 76 variables (corresponding to the features observed) and 128 buildings.

4.2 Application of DMTs

Regarding case study 2, the variable correlation analysis identified more than 60 strong correlations between the 76 variables composing the initial database, which allowed for the creation of a simplified database composed of 17 independent variables and 128 buildings (table 4).

Table 4. List of 17 independent variables considered for case study 2

Group of features	Variable code	Description
A: Geometry	V01	Number of floors: 1
	V04	% of openings in main facade < 10%
	V07	% of openings in main facade: non-applicable
	V09	Roof shape: hip
	V11	Roof shape: other
	V12	Roof shape: non-applicable
	V16	Number of roof slopes: 4
	V18	Geometry of main facade: regular
	V20	Opening alignment: vertically aligned with the wall
	V22	Opening alignment: non-applicable
B: Functionality and habitability	V23	Floor-plan shape: rectangular
	V35	Solar orientation of main facade: north
	V45	Main entrance: through other buildings
C: Technology and materials	V53	Window framing material: wood
	V56	Facade coating material: un-coated stone masonry
	V59	Facade coating material: clay bricks
	V69	Roof coating: clay tiles, barrel type

Then, the calculation of Mahalanobis distance and respective probability highlighted the existence of 14 outliers with probability values under 0.001 (Table 5). When analysed in detail, these buildings were found to have a set of features that are highly discrepant from the rest of the sample, such as the inclusion of modern and new materials, or profound alterations in geometry, volume, or functionality.

Table 5. Identification of outliers through the calculation of Mahalanobis distance and respective probability

Building code	Mahalanobis distance (D2)	D2 probability
CAV-010	2.747	0.99996
CAV-012	2.747	0.99996
CAV-014	2.747	0.99996
BAJ-003	2.747	0.99996
BAJ-018	2.747	0.99996
(intermediate values omitted)		
BAJ-021	44.029	**< 0.001**
MED-010	45.269	**< 0.001**
BAJ-015	47.317	**< 0.001**

BAJ-002	47.887	**< 0.001**
MDD-002	51.917	**< 0.001**
MED-061	53.948	**< 0.001**
PZM-008	126.008	**< 0.001**
BAJ-014	126.008	**< 0.001**
NRT-034	126.008	**< 0.001**
BAJ-017	126.008	**< 0.001**
BAJ-024	73.054	**< 0.001**
BAJ-030	73.054	**< 0.001**
MED-015	65.100	**< 0.001**
MED-045	65.100	**< 0.001**

Besides the calculation of Mahalanobis distance probability, the cluster dendrogram obtained from the hierarchical algorithm was also analysed, allowing together for the identification of one other building isolated from the agglomerative process (Figure 11). Overall, the identification of outliers allowed for obtaining a final sample of 113 buildings.

The visual inspection of the dendrogram obtained from the hierarchical clustering indicated an optimal number of 2 clusters (K), which allowed us to perform the non-hierarchical process (K-means) based on this input. Table 6 presents the probability of occurrence of a group of selected characteristics (corresponding to different variables of the database) in each cluster, as well as in the total sample. The probability of occurrence of each feature in each cluster was calculated on a scale that ranges from 0 (no building presents the feature) to 1 (all buildings present the feature) (Mouraz et al., 2022). This way, it is possible to compare the nature of each cluster with the total sample regarding a group of characteristics. The colour scale in Table 6 is used to better visualising the values of probability ranging from zero (red) to one (green).

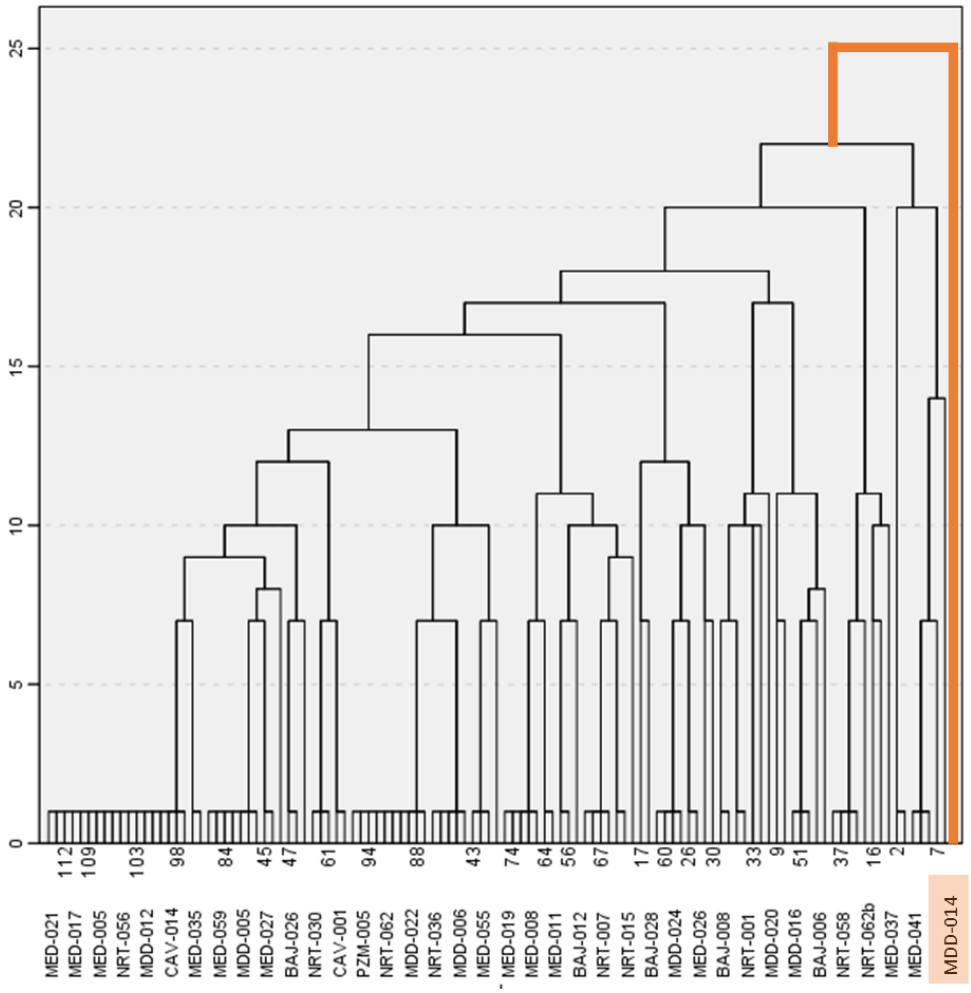


Figure 11. Cluster dendrogram obtained through hierarchical clustering algorithm for case study 2 and identification of outlier.

Table 6. Probability of occurrence of a set of variables in the final clustering solution for case study 2

	Features								
	V01 Number of floors: 1	**V04 % openings in main facade < 10%**	**V12 Roof shape: non-applicable**	**V20 Openings alignment: vertically alignedwith the wallface**	**V23 Floor-plan shape: rectangular**	**V35 Solar orientation of main facade: north**	**V53 Window framing material: wood**	**V56 façade coating material: uncoated stone masonry**	**V69 Roof coating: clay tiles. barrel type**
Total sample (113 buildings)	0.372	0.221	0.071	0.097	0.478	0.159	0.124	0.894	0.885
Cluster 1 (54 buildings)	0.296	0.167	0.019	0.074	1.000	0.148	0.148	0.944	0.926
Cluster 2 (59 buildings)	0.441	0.271	0.119	0.119	0.000	0.169	0.102	0.847	0.847

Contrary to case study 1 , where clusters presented meaningful differences regarding a significant set of variables (synthesized in Table 3), Table 6 points that the two clusters obtained for case study 2 don't have significant differences in the selected characteristics, except in the variable "V23- Floor plan shape: rectangular", which appears to be the defining aspect that differs between clusters - all buildings in Cluster 1 present this feature (since the probability of occurrence is 1), and no building in Cluster 2 presents this feature (since the probability of occurrence is 0).

Apart from this feature, the remaining variables present a similar probability of occurrence in both clusters, thus not being clear if any clusters are directly associated with buildings presenting characteristics closer to the vernacular matrix.

4.3 GIS visualisation

Finally, the visual representation of the clustering distribution was obtained through GIS (Figure 12).

Figure 12. Mapping of the final clustering solution found for case study 2.

It is possible to observe that buildings belonging to Cluster 2 are mainly located at the extreme ends of the settlement and along the main central street. Outliers are evenly

distributed along the settlement, and some present less conventional floor-plan shapes, which can be indicative of profound changes in the original geometry, validating discrepancies that may have contributed to their identification as outliers. Also, the differentiating characteristic between Clusters 1 and 2, the floor-plan shape, is also clearly visible on the map, with rectangular buildings identified as part of Cluster 1.

5. Discussion

5.1 Characterisation of case studies

Through this research, it was possible to characterise two rural settlements with vernacular construction. Despite being in different countries with singular territorial characteristics, both cases have suffered an ongoing process of depopulation and abandonment that reflects in the occupation of settlements. Both cases share a general context of abandonment, lack of permanent inhabitants, and poor state of conservation of buildings.

Both cases have intrinsic differences in urban and construction characteristics, mainly considering local conditions that influence materials used in vernacular constructions. Nonetheless, despite heterogeneities in materials such as the typical stone, both cases share barrel-type clay tiles as the most common type of roof coating.

The exterior survey carried out led, in both cases, to the creation of databases with a high number of variables directly related to the variety of features observed in each case. These variables were grouped into three sets of features: geometric, functionality and habitability, and technology and materials. Nonetheless, databases differed significantly in the number of buildings considered (case study 1 comprises 27 buildings, whilst case study 2 includes 128).

5.2 Application of DMTs

The first DMT applied concerned the reduction of data, and in both cases the results observed were significant, with a reduction of almost 57% of the variables in case study 1 and 78% in case study 2. The high number of correlated variables in a set of features initially perceived as independent shows the importance of this DMT in highlighting less obvious relations between features.

Then, for both cases, the outlier identification allowed for the selection of highly discrepant buildings. The innovative approach in which two methods for the identification of outliers is highlighted since it allowed for the identification of a higher number of outliers in case study 2. The calculation of the Mahalanobis distance only provided valid results in case study 2, probably due to the larger dimension of the sample and, thus, to a higher variability between buildings and corresponding features.

Cluster analyses were then conducted to reach an optimised solution by adopting a two-step process combining hierarchical and non-hierarchical clustering algorithms. For case study 1, characterised by fewer buildings, a higher number of clusters was obtained (K=4) compared to case study 2 (K=2). This result is coherent with the research works presented in Table 1 and the expectable conclusion that a higher number of buildings does not necessarily leads to more clusters since the number of clusters depends on the homogeneity of the dataset.

For case study 1, it was identified one cluster that includes 47% of buildings in the dataset. This cluster encompasses the buildings that are still close to the original vernacular matrix, including, as distinctive features, uncoated schist masonry apparatus, clay barrel-type tiles, or wooden window elements. The remaining clusters were related to summer houses (Clusters 2 and 4) and buildings under renovation works (Cluster 3), which is coherent with the occupation of the observed buildings in this mountain

settlement. This indicates that occupation seems to have been the main differentiating aspect in defining clusters.

Regarding case study 2, excluding previously identified outliers, two clusters were disclosed in the remaining sample. When observing a set of features, no significant discrepancies were found between the two clusters, except for the floor-plan shape, which appears to have been the differentiator feature. However, the predominance of characteristics such as uncoated limestone masonry, wooden elements, and clay barrel-type tiles, validate the assumption regarding the existence of buildings where vernacular characteristics prevail. Even though more studies are required to locate these buildings within the clusters, detecting a substantial set of discrepant buildings as outliers is particularly useful in the developed research, highlighting the importance of the combined approach followed.

The conducted research allowed to obtain databases with a higher number of variables compared to other research works, which may pose a limitation towards the goal of clearly identifying typologies of buildings within the case studies. Despite validating the existence of buildings with characteristics closer to the vernacular matrix, reorganising relevant information regarding the characterisation of buildings and adapting databases to cover key aspects while reducing the number of variables may pose a more effective input to future cluster analyses that can disclose a more efficient output regarding the presence of vernacular buildings. However, such does not come without some additional challenges. The first and a significant one concerns the definition of the set of variables to be included in the database, given that only the right set of variables can lead to more accurate results in identifying these buildings.

Nevertheless, the proposed methodology can be applied to other contexts based on the development of specific databases that suffice to characterise each reality.

5.3 GIS visualisation

Finally, GIS maps made it possible to graphically visualise the obtained clustering results. For case study 1, certain features that characterise clusters, such as floor-plan shapes and north-orientated facades, are clearly visible. However, the high number of clusters hinders the interpretation of eventual spatial patterns and precludes further conclusions.

Regarding case study 2, the lack of significant differences between buildings composing Clusters 1 and 2 contributes to the lack of further conclusions surrounding the location of buildings with vernacular characteristics. However, important conclusions are drawn considering the significant number of outliers identified and features observed in the map, such as irregular floor plan shapes.

6. Conclusions

Despite their broad scope of use, research on the application of DMTs to different-sized samples of buildings is limited. Also, the use of vernacular constructions in rural territories as case studies remains rare, especially focusing on the pertinence of future rehabilitation actions.

This paper presents a methodology for characterising two rural settlements with vernacular construction, in Portugal and Spain, including a discussion on the application of DMTs. This work intends to raise awareness of undocumented built heritage, apply cluster analysis to identify similar groups of buildings, validate the existence of buildings with characteristics closer to the vernacular matrix, and examine results obtained for different objects.

The presented approach led to valid results on the efficiency of applying DMTs such as data cleaning, outlier detection and cluster analysis to the study objects.

Preliminary results validated the existence of buildings presenting a set of features closer to the vernacular matrix through cluster analyses. Even though thorough approaches are needed regarding the detailed identification of these buildings, the application of other DMTs, such as outlier detection, was also particularly useful in identifying sets of dissonant buildings.

Future research will focus on three groups of action. First, we will improve the proposed methodology considering an adaptation of database and the collected information to limit the number of variables considered, developing that way a narrower approach focused on key aspects that may lead to a clear identification of vernacular buildings in both study cases and significantly reduce the number of variables considered, which may pose a challenge due to the need for a set of aspects covering aspects that can fully characterise these realities. Secondly, we will compare the accuracy of results obtained through other clustering techniques in identifying typologies of buildings in the case studies considering, namely using statistical-based approaches in establishing the optimal number of clusters, and analysing the accuracy of these results towards the goals of identifying a group of buildings closer to the vernacular matrix. Thirdly, it is essential to conduct a qualitative assessment of these territories analysing complementary perspectives that favour their holistic characterisation encompassing historical, social or geographical perspectives. Narrowing the knowledge gap surrounding these settlements and their constructions contributes towards a qualified input in future territorial policies and leads the way towards their sustainable conservation and improvement.

This research lays the foundations for future rehabilitation and conservation actions in vernacular buildings in rural settlements. Among other aspects, it contributes

to making more informed decisions regarding the characterisation of these constructions and planning for future improvements in their performance.

Acknowledgements

This research was funded by the European Union through the European Social Found and the Portuguese Foundation for Science and Technology (FCT) through the grant with reference 2021.07322.BD.

Disclosure statement

The authors report there are no competing interests to declare.

References

Afaifia, Marwa, Kahina Amal Djiar, Nguyen Bich-Ngoc, and Jacques Teller. 2021. "An Energy Consumption Model for the Algerian Residential Building's Stock, Based on a Triangular Approach: Geographic Information System (GIS), Regression Analysis and Hierarchical Cluster Analysis." *Sustainable Cities and Society* 74 (November): 103191. https://doi.org/10.1016/J.SCS.2021.103191.

Alexiadis, Stilianos. 2017. "Territorial Cohesion and Prospects for Sustainable Development: A Co-Integration Analysis." *Habitat International* 68: 75–83. https://doi.org/10.1016/j.habitatint.2017.03.001.

Arambula Lara, Rigoberto, Francesca Cappelletti, Piercarlo Romagnoni, Andrea Gasparella, Arambula Lara, and Rigoberto Arambula. 2014. "Selection of Representative Buildings through Preliminary Cluster Analysis." *International High Performance Buildings Conference*. http://docs.lib.purdue.edu/ihpbc/137.

Bussab, W. de O., Miazaki, E. S., & Andrade, D. F. de. (1990). Introdução à análise de agrupamentos. São Paulo: IME-USP

Carvalho, P. 2009. Património Construído e Desenvolvimento Em Áreas de Montanha. O Exemplo Da Serra Da Lousã. 2nd edition. Lousã: Câmara Municipal da Lousã.

Cheng, Geng, Zao Li, Shuting Xia, Mingfei Gao, Maosheng Ye, and Tingting Shi. 2023. "Research on the Spatial Sequence of Building Facades in Huizhou Regional Traditional Villages." *Buildings* 13 (1): 1–31. https://doi.org/10.3390/buildings13010174.

Elert, Kerstin, Eva García Baños, Aurelia Ibañez Velasco, and Pedro Bel-Anzué. 2021. "Traditional Roofing with Sandstone Slabs: Implications for the Safeguarding of Vernacular Architecture." *Journal of Building Engineering* 33 (January): 101857. https://doi.org/10.1016/j.jobe.2020.101857.

European Commission (n.d.). "Mountains, Islands and Sparsely Populated Areas". Accessed on 14th June 2023. https://ec.europa.eu/regional_policy/en/policy/themes/sparsely-populated-areas/#3

Gulagiz, Fidan Kaya, and Sahin Suhap. 2017. "Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms." *International Journal of Computer Engineering and Information Technology* 9 (1): 6–14. www.ijceit.org.

Hair, J. F., W.C. Black, B.J. Babin, and J.E. Anderson. 2007. *Multivariate Data Analysis*. 7th edition. Pearson.

He, Yi, Yingnan Chu, Yehao Song, Mengjia Liu, Shaohang Shi, and Xinxing Chen. 2022. "Analysis of Design Strategy of Energy Efficient Buildings Based on Databases by Using Data Mining and Statistical Metrics Approach." *Energy and Buildings* 258 (March): 111811. https://doi.org/10.1016/J.ENBUILD.2021.111811.

Iglesias Vazquez, Félix, and Wolfgang Kastner. 2013. "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns." *Energies* 6 (2): 579–97. https://doi.org/10.3390/en6020579.

Javatpoin. "Data cleaning in data mining". Accessed on 12th may 2023. https://www.javatpoint.com/data-cleaning-in-data-mining

Jiang, Wei, Yang Jin, Gongliang Liu, Qing Li, and Dong Li. 2023. "Passive Nearly Zero Energy Retrofits of Rammed Earth Rural Residential Buildings Based on Energy Efficiency and Cost-Effectiveness Analysis." *Renewable and Sustainable Energy Reviews* 180 (July): 113300. https://doi.org/10.1016/J.RSER.2023.113300.

Li, Xinyi, Runming Yao, Meng Liu, Vincenzo Costanzo, Wei Yu, Wenbo Wang, Alan Short, and Baizhan Li. 2018. "Developing Urban Residential Reference Buildings

Using Clustering Analysis of Satellite Images." *Energy and Buildings* 169 (June): 417–29. https://doi.org/10.1016/J.ENBUILD.2018.03.064.

Li, Qing, Hao Hu, Lingyong Ma, Zhiguo Wang, Müslüm Arıcı, Dong Li, Dan Luo, Jiaojiao Jia, Wei Jiang, and Hanbing Qi. 2022. "Evaluation of Energy-Saving Retrofits for Sunspace of Rural Residential Buildings Based on Orthogonal Experiment and Entropy Weight Method." *Energy for Sustainable Development* 70 (October): 569–80. https://doi.org/10.1016/J.ESD.2022.09.007.

Mayrhofer, Marcus, and Peter Filzmoser. 2023. "Multivariate Outlier Explanations Using Shapley Values and Mahalanobis Distances." *Econometrics and Statistics*, April. https://doi.org/10.1016/J.ECOSTA.2023.04.003.

Mouraz, Catarina P., Ricardo M.S.F. Almeida, and J. Mendes Silva. 2022a. "Combining Cluster Analysis and GIS Maps to Characterise Building Stock: Case Study in the Historical City Centre of Viseu, Portugal." *Journal of Building Engineering* 58 (April): 104949. https://doi.org/10.1016/j.jobe.2022.104949.

Mouraz, Catarina P., J. Mendes Silva, and Tiago Miguel Ferreira. 2022b. "Portuguese Vernacular Construction and Its Sustainable Rehabilitation Challenges: The Schist Villages, Lousã." In *REHABEND 2022 Congress: Construction Pathology, Rehabilitation Technology and Heritage Management*, edited by Haydee Blanco, Iosbel Boffill, and Ignacio Lombillo, 1402–10. Granada: University of Cantabria. https://bit.ly/42A0rFp

Mouraz, Catarina P, Tiago Miguel Ferreira, and J. Mendes Silva. 2023. "Building Rehabilitation, Sustainable Development, and Rural Settlements: A Contribution to the State of the Art." *Environment, Development and Sustainability*. https://doi.org/10.1007/s10668-023-03664-5.Município da Lousã, Gabinete Técnico Florestal. *Plano Municipal de defesa da floresta contra incêndios 2020-2029- Caderno I: Diagnóstico.* Lousã, 2020. Accessed September 15, 2023. https://cm-lousa.pt/wp-content/uploads/2017/04/CADERNO-I.pdf

Nettleton, David. 2014. "Selection of Variables and Factor Derivation." *Commercial Data Mining*, January, 79–104. https://doi.org/10.1016/B978-0-12-416602-8.00006-6.

Nguyen, Anh Tuan, Nguyen Song Ha Truong, David Rockwood, and Anh Dung Tran Le. 2019. "Studies on Sustainable Features of Vernacular Architecture in Different Regions across the World: A Comprehensive Synthesis and Evaluation." *Frontiers*

of Architectural Research 8 (4): 535–48. https://doi.org/10.1016/j.foar.2019.07.006.

Obilor, Esezi, Amadi, Eric. 2018. "Test for Significance of Pearson's Correlation Coefficient (r)". *International Journal of Innovative Mathematics, Statistics & Energy Policies* 6(1): 11-23, Jan-Mar. https://www.researchgate.net/publication/323522779_Test_for_Significance_of_Pearson's_Correlation_Coefficient

Parracha, João Luís, José Lima, Maria Teresa Freire, Micael Ferreira, and Paulina Faria. 2021. "Vernacular Earthen Buildings from Leiria, Portugal – Architectural Survey towards Their Conservation and Retrofitting." *Journal of Building Engineering* 35 (March): 102115. https://doi.org/10.1016/J.JOBE.2020.102115.

Paniagua, Angel. 2008. "The Environmental Dimension in the Constitution of New Social Groups in an Extremely Depopulated Rural Area of Spain (Soria)." *Land Use Policy* 25 (1): 17–29. https://doi.org/10.1016/j.landusepol.2007.02.001.Patteeuw, D., Gregor P. Henze, Alessia Arteconi, Charles D. Corbin, and Lieve Helsen. 2019. "Clustering a Building Stock towards Representative Buildings in the Context of Air-Conditioning Electricity Demand Flexibility." *Journal of Building Performance Simulation* 12:1: 56–67. https://doi.org/10.1080/19401493.2018.1470202.

Pistore, Lorenza, Giovanni Pernigotto, Francesca Cappelletti, Andrea Gasparella, and Piercarlo Romagnoni. 2019. "A Stepwise Approach Integrating Feature Selection, Regression Techniques and Cluster Analysis to Identify Primary Retrofit Interventions on Large Stocks of Buildings." *Sustainable Cities and Society* 47 (May): 101438. https://doi.org/10.1016/j.scs.2019.101438.

Ray, Biswarup, Soulib Ghosh, Shameem Ahmed, Ram Sarkar, and Mita Nasipuri. 2022. "Outlier Detection Using an Ensemble of Clustering Algorithms." *Multimedia Tools and Applications* 81 (2): 2681–2709. https://doi.org/10.1007/s11042-021-11671-9.

Rosti, A., M. Rota, and A. Penna. 2022. "An Empirical Seismic Vulnerability Model." *Bulletin of Earthquake Engineering* 20 (8): 4147–73. https://doi.org/10.1007/s10518-022-01374-3.

Sambandam, R. 2003. “Cluster Analysis Gets Complicated.” *Marketing Research* 15 (1): 16–21. https://trcmarketresearch.com/whitepaper/cluster-analysis-gets-complicated/.

Schaefer, Aline, and Enedir Ghisi. 2016. “Method for Obtaining Reference Buildings.” *Energy and Buildings* 128 (September): 660–72. https://doi.org/10.1016/J.ENBUILD.2016.07.001.

Shan, Xiaofang, Qinli Deng, Zheng Tang, Zhi Wu, and Wei Wang. 2022. “An Integrated Data Mining-Based Approach to Identify Key Building and Urban Features of Different Energy Usage Levels.” *Sustainable Cities and Society* 77 (August 2021): 103576. https://doi.org/10.1016/j.scs.2021.103576.

Szabó, Simon, Marco Francesco Funari, and Paulo B. Lourenço. 2023. “Masonry Patterns’ Influence on the Damage Assessment of URM Walls: Current and Future Trends.” *Developments in the Built Environment* 13 (November 2022). https://doi.org/10.1016/j.dibe.2023.100119.

Tardioli, Giovanni, Ruth Kerrigan, Mike Oates, James O’Donnell, and Donal P. Finn. 2018. “Identification of Representative Buildings and Building Groups in Urban Datasets Using a Novel Pre-Processing, Classification, Clustering and Predictive Modelling Approach.” *Building and Environment* 140 (August): 90–106. https://doi.org/10.1016/j.buildenv.2018.05.035.

The World Bank, “Rural population (% of total population)”. Accessed May 17, 2023. https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS

Vicente, Romeu, Tiago Miguel Ferreira, and J. A. Raimundo Mendes da Silva. 2015. “Supporting Urban Regeneration and Building Refurbishment. Strategies for Building Appraisal and Inspection of Old Building Stock in City Centres.” *Journal of Cultural Heritage* 16 (1): 1–14. https://doi.org/10.1016/j.culher.2014.03.004.

Wikipedia. “Lousã”. *Wikipedia*. Accessed September 15, 2023a. https://pt.wikipedia.org/wiki/Lous%C3%A3.

Wikipedia. “Rello”. *Wikipedia.* Accessed September 15, 2023b. https://es.wikipedia.org/wiki/Rello

Yan, Qingli, Jianfeng Chen, and Lieven De Strycker. 2018. "An Outlier Detection Method Based on Mahalanobis Distance for Source Localization" *Sensors* 18, no. 7: 2186. https://doi.org/10.3390/s18072186

Yim, Odilia, and Kylee T. Ramdeen. 2015. “Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data.” The

Quantitative Methods for Psychology 11 (1): 8–21. https://doi.org/10.20982/tqmp.11.1.p008.

Yin, Ximing, Jin Chen, and Jizhen Li. 2019. “Rural Innovation System: Revitalize the Countryside for a Sustainable Development.” *Journal of Rural Studies*, October. https://doi.org/10.1016/j.jrurstud.2019.10.014.

Zhang, Chuan, Liwei Cao, and Alessandro Romagnoli. 2018. “On the Feature Engineering of Building Energy Data Mining.” *Sustainable Cities and Society* 39 (May): 508–18. https://doi.org/10.1016/J.SCS.2018.02.016.

Zhao, Yang, Chaobo Zhang, Yiwen Zhang, Zihao Wang, and Junyang Li. 2020. “A Review of Data Mining Technologies in Building Energy Systems: Load Prediction, Pattern Identification, Fault Detection and Diagnosis.” *Energy and Built Environment* 1 (2): 149–64. https://doi.org/10.1016/J.ENBENV.2019.11.003.