

A data-driven energy platform: from energy performance certificates to human-readable knowledge through dynamic high-resolution geospatial maps

*Original*

A data-driven energy platform: from energy performance certificates to human-readable knowledge through dynamic high-resolution geospatial maps / Cerquitelli, T.; Corso, E. D.; Proto, S.; Bethaz, P.; Mazzearelli, D.; Capozzoli, A.; Baralis, E.; Mellia, M.; Casagrande, S.; Tamburini, M.. - In: ELECTRONICS. - ISSN 2079-9292. - 9:12(2020), pp. 1-26. [10.3390/electronics9122132]

*Availability:*

This version is available at: 11583/2866059 since: 2021-01-23T07:37:50Z

*Publisher:*

MDPI

*Published*

DOI:10.3390/electronics9122132

*Terms of use:*





This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

## Article

# A Data-Driven Energy Platform: From Energy Performance Certificates to Human-Readable Knowledge through Dynamic High-Resolution Geospatial Maps

Tania Cerquitelli <sup>1,\*</sup> , Evelina Di Corso <sup>1</sup> , Stefano Proto <sup>1</sup>, Paolo Bethaz <sup>1</sup> , Daniele Mazzearelli <sup>2</sup>, Alfonso Capozzoli <sup>2</sup> , Elena Baralis <sup>1</sup>, Marco Mellia <sup>3</sup>, Silvia Casagrande <sup>4</sup> and Martina Tamburini <sup>4</sup>

<sup>1</sup> Department of Control and Computer Engineering, Politecnico di Torino, 10129 Torino, Italy; evelina.dicorso@polito.it (E.D.C.); stefano.proto@polito.it (S.P.); paolo.bethaz@polito.it (P.B.); elena.baralis@polito.it (E.B.)

<sup>2</sup> Department of Energy “Galileo Ferraris”, Politecnico di Torino, 10129 Torino, Italy; daniele.mazzearelli@polito.it (D.M.); alfonso.capozzoli@polito.it (A.C.)

<sup>3</sup> Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Torino, Italy; marco.mellia@polito.it

<sup>4</sup> Edison S.p.a., 10129 Torino, Italy; silvia.casagrande@edison.it (S.C.); martina.tamburini@edison.it (M.T.)

\* Correspondence: tania.cerquitelli@polito.it; Tel.: +39-011-090-7178

Received: 20 November 2020; Accepted: 8 December 2020; Published: 12 December 2020



**Abstract:** The energy performance certificate (EPC) is a document that certifies the average annual energy consumption of a building in standard conditions and allows it to be classified within a so-called energy class. In a period such as this, when greenhouse gas emissions are of considerable importance and where the objective is to improve energy security and reduce energy costs in our cities, energy certification has a key role to play. The proposed work aims to model and characterize residential buildings’ energy efficiency by exploring heterogeneous, geo-referenced data with different spatial and temporal granularity. The paper presents TUCANA (Turin Certificates ANALysis), an innovative data mining engine able to cover the whole analytics workflow for the analysis of the energy performance certificates, including cluster analysis and a model generalization step based on a novel spatial constrained K-NN, able to automatically characterize a broad set of buildings distributed across a major city and predict different energy-related features for new unseen buildings. The energy certificates analyzed in this work have been issued by the Piedmont Region (a northwest region of Italy) through open data. The results obtained on a large dataset are displayed in novel, dynamic, and interactive geospatial maps that can be consulted on a web application integrated into the system. The visualization tool provides transparent and human-readable knowledge to various stakeholders, thus supporting the decision-making process.

**Keywords:** data exploration; data visualization; energy performance certificates; energy maps; spatial constrained K-NN

## 1. Introduction

According to the U.S. Department of Energy, in industrialized countries, more than 40% of total energy is consumed in buildings [1]. It is therefore necessary to implement a process of environmental, energy, and economic regeneration of the whole sector. In several countries, regulatory bodies took action to reduce energy waste by encouraging the use of renewable energy sources and the design of

energy-efficient buildings [2]. The energy certification of a building indicates a collection and analysis of the energy performance, in order to identify the weakness and causes of possible vulnerability.

This information is contained in Energy Performance Certificates (EPCs), which provide indications about energy performance and thermo-physical and geometrical related properties of a building. The exploration of these certificates is a key point to correctly assess building energy related performance within a city or region and the information extracted from this data is fundamental to decide how to implement energy-saving services or refurbishment strategies in the analyzed territory. The assessment of energy performance in buildings has found great interest in the last few years from the multidisciplinary research community. From a design perspective, the latent main relationships between the building features and the future energy performance are key research challenges [3]. However, various research issues, exploiting EPC as input data, could lead to different insights and effectively support the decision-making process.

This paper presents TUCANA (Turin Certificates ANALysis), a data-driven methodology for the energy related characterization of buildings in the city of Turin, through the use of data analytics, data mining and machine learning techniques applied on geometric, thermo-physical, and system based features gathered from energy performance certificates (EPCs). To do this, TUCANA has been designed and developed, providing different data analytics capabilities:

- *A multivariate characterization (energy modeling) of buildings based on unsupervised techniques (i.e., cluster analysis) to identify cohesive groups of buildings, similar in terms of geometry, thermo-physical and system efficiency related features.*
- *Prediction of different energy-related features for new unseen buildings.* It relies on a novel variant of the K-NN algorithm [4] that also includes the selection of the set of features to be analyzed through a spatial metrics. This ability enriches TUCANA with predicting capability whose results allow providing a more-detailed energy mapping including those buildings for which there is no EPC available, or the EPC is only partially complete or includes inconsistent values of influencing variables.
- *Visualization of the analytics results through novel, interactive and geolocalized maps, with different levels of granularity and complexity, addressing the needs and expertise of different stakeholders.* Several dynamic and geolocalized maps have been integrated, with different levels of complexity that make them suitable for different stakeholders, making the knowledge as understandable as possible to stakeholders who may have different skills and interests. This feature could easily support the decision-making process.
- *A user-friendly accessible tool to easily exploit the analytics insights.* TUCANA includes a web application able to explore the available data and to make the extracted knowledge easily consultable and used by the various end users. The purpose of this tool is to show to different stakeholders only the the results of their interest in a user-friendly way. Since statistical representations are understandable especially by domain experts and can be difficult to interpret by general users, the results are reported through more readable maps, tailored to different end-users.

The paper is organized as follows. Section 2 describes the state of the art works related to the proposed context. Section 3 details the proposed methodology, introducing the main building blocks of the framework. Section 4 describes the results obtained by applying TUCANA in a real use case, analyzing energy data of the Piedmont region. Finally, Section 5 discusses open issues and presents the future development of this work.

## 2. Literature Review

The data-driven methodologies applied on energy-related data are proposed in a large body of literature. Studies can be categorized based on the main research issue addressed into four main classes: (1) data characterization and exploration [5–9]; (2) data management and architectures [7,10]; (3) predictive analytics [11–16]; and (iv) geospatial data visualization [7,17–20].

Different kinds of data can be exploited to discover latent but useful relationships from energy-related data. In Kaden and Kolbe [5], the building's energy demand is estimated by analyzing the thermo-physical properties of buildings in addition to the year of construction, while Howard et al. [8] focused on the estimation of the building sector energy end-use intensity (kWh/m<sup>2</sup> floor area), assuming that such end-use is primarily dependent on building function. A similar energy mapping has been done in Evans et al. [21], which offers a 3D model for buildings in England and Wales, also representing the pattern of activities on different floors within buildings. Similar to the work in Di Corso et al. [6], Dall'O' et al. [22], the data used in our study are all open-data, as EPCs are available as such. The use of this kind of data is due to their great availability and the rich information content within them. However, in this paper, we address a completely different research issue (i.e., providing dynamic high-resolution geospatial maps) and propose an innovative data-driven solution with a significant new contributions (e.g., predictive Modeling and Visualization of energy maps able to represent interesting knowledge items) over the work proposed in Di Corso et al. [6]. Dall'O' et al. [22] discussed analyzing a real collection of EPC to estimate the energy performance, easily identify anomalies in the available database, and assess the potential benefit of energy retrofit in existing buildings.

The analysis of energy-related data has received great attention in various countries [23]. Hjortling et al. [24] combined the EPC data with energy bills for many buildings located in Sweden. The results demonstrate that the actual energy consumption is usually higher than the one calculated and available on EPC data. The research work presented in Xiao et al. [25] discusses the analysis performed on a large set of buildings located in China, showing that the statistical distribution characteristics of buildings in China were different from those in Japan and the US. The study of buildings' energy performance in Greece and Spain is discussed in Dascalaki et al. [26] and Gangolells et al. [27], respectively, with the final aim to discuss the energy consumption baseline.

Relevant design aspects related to the deployment of an architectural framework for energy-efficiency are discussed in Beloglazov et al. [10]. A Green Cloud computing approach able to reduce the environmental impact is presented. On the other hand, the design and the development of a tool to estimate the energy performance for residential buildings at city scale is discussed in Agugiaro [7]. Heterogeneous datasets (cadastral data, statistical data, etc.) were harmonized and integrated with the final aim of classifying residential buildings into distinct building types according to the criteria defined for Italy in the European project "Tabula" [9].

Much research has been devoted to designing data-driven methodologies to estimate energy-related parameters. As the highest energy consumption within an industrialized city comes from buildings, in Zhao and Magoulès [12], the authors tested different predictive models to estimate the building energy consumption, while Luo et al. [13] developed an online energy consumption and production prediction system, in order to provide energy prediction in a business district in Shanghai. Artificial neural networks (ANNs) are used in Kalogirou and Bojic [14] to forecast the energy consumption of a passive solar building, while Dong et al. [15] exploited Support Vector Machines (SVM) to forecast building energy consumption in the tropical region. A parallel research effort has been devoted to estimating the energy consumption of hosts and networks within virtualized and ordinary office [16]. In addition to what has been done in the studies described above, our methodology offers a strategy based on a clustering algorithm capable of discovering groups of EPCs characterized by similar geometric, thermo-physical, and system related features; a novel classification algorithm based on K-NN to estimate missing or inconsistent energy-related input features; and a visualization part that allows the user to see the results obtained in the previous points on maps with different granularity.

Some research efforts have been also devoted to considering the geographical location of the buildings in the analysis of building energy consumption. Strzalka et al. [17] presented a 3D model coupled with Geographical Information System (GIS) to visualize heating energy consumption of buildings. In Heiple and Sailor [18], the authors explore annual building energy simulations for a few

city-specific prototypical buildings and a few geospatial data to estimate hourly and seasonal energy consumption profiles of buildings. Geographic data are then often visualized on comfortable visual maps, e.g. Caputo et al. [28] visualized on a city-level map the energy consumption disaggregated on the final use (i.e., cooking and heating). A 3D map on Google Earth at district-level was proposed by Agugiaro [7] to visualize some interesting statistics on energy consumption, while Sicilia et al. [19] computed descriptive statistics over Energy Performance Certificates are visualized on a map. I. [20] discussed the key aspects and the main objectives that drove the London Building Stock Model's development. The proposed energy map shows the energy performance information related to a sample of buildings located in London. The proposed visualization's main objective is to identify the worst-performing buildings that probably correspond to the homes that mainly need to be refurbished to improve their fuel poverty.

Some preliminary results of TUCANA are presented in Cerquitelli et al. [29,30]. While a very limited graphical visualization of EPCs on maps is discussed in Cerquitelli et al. [29], a two-level clustering strategy is presented in Cerquitelli et al. [30] to characterize a collection of EPCs. However, the study presented here significantly improves our previous works, proposing a data-driven strategy capable of automatically grouping together EPCs with similar thermo-physical properties. Moreover, these results can be visualized graphically on maps at a different level of resolution. In particular, TUCANA: (a) characterizes each cluster, jointly considering the effect of several variables that are able to influence the energy performance of a building; (b) generalizes the knowledge obtained during the clustering phase, being able to predict input influencing features, thanks to a variant of K-NN based on spatial constraints; and (c) graphically shows the extracted knowledge, through the use of navigable maps.

### 3. TUCANA Framework

This paper presents an innovative data mining engine, named TUCANA (Turin Certificates ANALysis), covering the whole analytics workflow for the analysis of the energy performance certificates. TUCANA analyzes EPC collections through a two-level methodology based on a cluster analysis and a model generalization step to automatically extract patterns able to perform well even on new unseen data.

This joint approach allows a specific characterization of the EPCs. Specifically, the clustering phase is able to divide the input collection into homogeneous groups of dwellings with similar geometric, thermo-physical, and system efficiency related features. Then, each cluster is characterized using different statistical procedures which are able to highlight interesting patterns among the data. This procedure allows classifying the performance of each cluster according to their features. During the model generalization step, different models are built on existing data to generalize the extracted knowledge for future data. Generalization indicates the adaptability of the model even on new data, never seen before.

TUCANA exploits the K-means algorithm [31] to cluster EPCs, choosing the best K value thanks to an automatic methodology, while the generalization phase extracts models for the predictive analytics. TUCANA also includes a visualization phase to graphically represent the extracted knowledge from clustering, through the use of different geospatial maps. These maps represent a different level of information content, making the tool usable by different stakeholders (e.g., energy experts, public administration, and citizens) who are helped in their decision making process.

The main novelties of TUCANA are three-fold.

1. *Descriptive Modeling*: To automatically discover homogeneous groups of dwellings characterized by similar geometric, thermo-physical, and system related features. Groups are found using a clustering algorithm and a subsequent characterization of the discovered clusters, evaluating multiple variables that can influence the energy efficiency of a building.
2. *Predictive Modeling*: Since several buildings do not have their own EPC or have an EPC affected by missing or inconsistent values of input variables, this step allows generalizing the knowledge

extracted, increasing the number of available data and therefore including in the proposed methodology a larger number of buildings. This task is done using a spatial-constrained K-NN, a novel variant of K-NN that also includes the selection of the set to be analyzed with a spatial metric.

3. *Visualization* of energy maps able to represent the various clusters, considering at the same time the effect of multiple influencing variables. The different levels of resolution that the various kind of maps offer make the visualization suitable for multiple stakeholders.

Figure 1 shows the overall architecture of TUCANA and represents the whole data analytics workflow. It includes three main blocks; (1) *Data pre-processing*; (2) *Descriptive modeling and exploitation*; and (3) *Predictive modeling*.

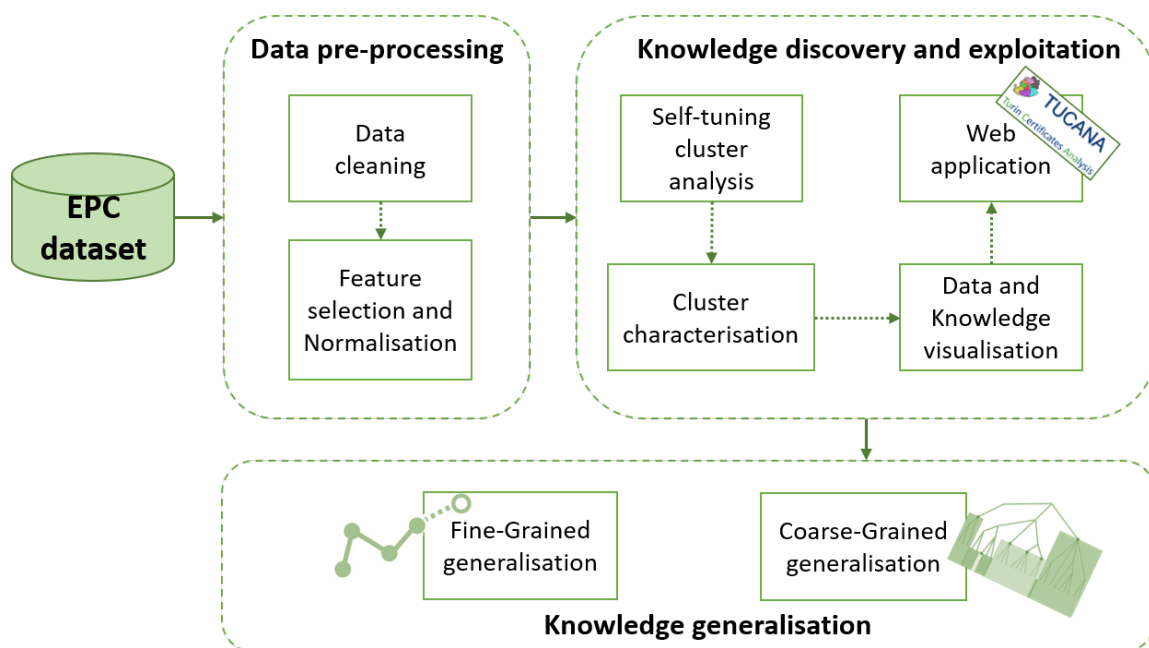


Figure 1. The TUCANA architecture.

### 3.1. Data Pre-Processing

Clearly, after the collection of EPCs data, the first step is the data pre-processing, in order to select only consistent data assuming values included in a range that is feasible from physics-based point of view for the following analysis. In particular, this phase can be divided into three main steps: (1) geospatial data cleaning; (2) outlier detection and removal; and (3) feature selection and normalization.

**Geospatial data cleaning.** This step checks that the geographical data (e.g., addresses, latitude, and longitude) contained in the EPCs are actually correct. Since the final aim of the methodology is to display the various buildings on a map, it is essential that there are no errors in the geographical information. As in Cerquitelli et al. [29], TUCANA compares the addresses available in the EPCs under analysis with those in the city's street database and evaluates their distance through the Levenshtein similarity, based on the computation of the Levenshtein distance, which is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into the other. If this similarity is higher than a given threshold (i.e., 0.95 in TUCANA framework), the comparison is considered successful, while, if the similarity is lower than the defined threshold, geocoding requests are sent using the Google Geocoding APIs [32] (<https://developers.google.com/maps/documentation/geocoding/intro>) to obtain the correct geospatial information (coordinates and addresses). At the beginning, this service allowed sending a limited number of free requests to the server; however, being a very precise and reliable service, over the years its services are no longer free.

**Outlier detection and removal.** It is not uncommon for open data to have several outliers, which are observations that largely deviate from the other data. Therefore, this step aims to identify them through both univariate and multivariate analysis coupled with a validation process conducted with a domain expert.

In particular, the univariate outlier analysis consists of three different stages:

1. As a first stage, outliers are identified with **generalized Extreme Studentized Deviate** (gESD) [33]. This technique requires as parameters the number of outliers ( $r$ ). This parameter is set by a domain expert and, given the upper bound value for the number of outliers, the gESD test essentially performs  $r$  separate tests: a test for one outlier, a test for two outliers, and so on up to  $r$  outliers.
2. Analysis of the **frequency data distribution** to identify outliers in the first percentile. TUCANA removes the points belonging to the first percentile, to eliminate inconsistent values with a low deviation from the rest of the observations.
3. Definition of **acceptability ranges** for each of the considered energy-related attribute, evaluated with the support of the energy expert, considering the physical significance of each variable.

Lastly, TUCANA identifies outliers through a multivariate analysis using the Density-Based Spatial Clustering of Application with Noise algorithm (DBSCAN) [34]. This cluster algorithm requires two input parameters (*minPoints* and *Epsilon*), which TUCANA tries to set in an automatic way exploiting the methodology presented in Ankerst et al. [35].

**Feature selection and normalization.** To consider only the most significant attributes among those present in the EPC, a correlation analysis between variables is performed. The correlation coefficient is computed through the Pearson correlation, defined as  $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ , where  $cov(X,Y)$  is the covariance between the two variables  $X$  and  $Y$ , and  $\sigma$  represents the standard deviation.  $\rho_{X,Y}$  can assume values between  $-1$  and  $+1$ , where  $+1$  corresponds to perfect positive linear correlation,  $0$  corresponds to no linear correlation, and  $-1$  corresponds to perfect negative linear correlation. If different variables are highly correlated, it becomes difficult to accurately partition out their individual impact on the analysis. Then, a max-min binormalization is used to eliminate redundant data and ensures that good quality clusters are generated, which can improve the efficiency of data mining algorithms. This stage becomes an essential step before clustering as distance is very sensitive to the changes in the differences.

### 3.2. Descriptive Modeling and Exploitation

This component entails the complex task of performing non-trivial processes with the final aim of identifying valid, novel, potentially useful and understandable patterns in the data under analysis. It includes several steps performed sequentially: (1) *Self-tuning cluster analysis*; (2) *Cluster characterization*; (3) *Data and Knowledge visualization*; and (4) *Web application*.

#### 3.2.1. Self-Tuning Cluster Analysis

To discover homogeneous groups of dwellings characterized by similar geometric, thermo-physical and system efficiency related features, TUCANA adopts the K-means clustering algorithm [31]. This is an unsupervised clustering algorithm able to find a  $K$  number (defined a priori) of clusters in a dataset. The algorithm is based on the concept of centroids (average of all features within the same cluster). These  $k$  centroids are initially chosen randomly, then each point is assigned to the nearest cluster, and for each iteration the position of the centroids is recalculated. Iterations end when the centroids no longer change (a point of convergence has been reached). To automatically choose the best  $K$  value to use in the algorithm, TUCANA analyzes the SSE (sum of squared error) trend [36] by increasing  $K$  and, as done in Satopaa et al. [37], it selects the value where the curvature is maximum, i.e. the *elbow* point which is furthest from the straight line between the first point and the last point of the SSE trend.

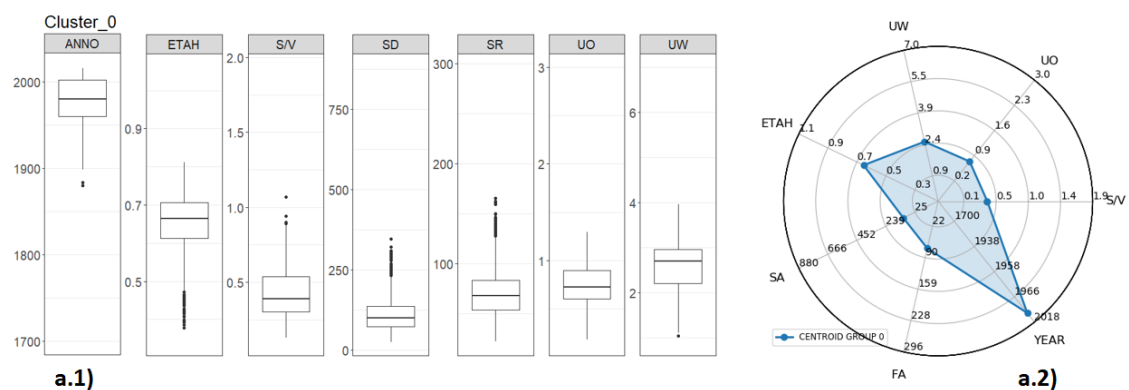
### 3.2.2. Cluster Characterization

Cluster analysis allows the discovery of groups of dwellings with similar properties; however, more human-readable awareness easily validated by a domain expert should be provided. To assess how effective the division into clusters has been, several graphic methods are integrated. The purpose of these graphs is to show users in an intuitive way how the identified clusters differ from each other, showing how the values of the various attributes are distributed within them. We speak of ‘goodness’ of clustering result according to how easily the different groups can be distinguished and characterized (i.e., if the distribution of attributes is clearly divided between the various clusters).

In the cluster characterization component, several different techniques are included in TUCANA to depict the goodness of the discovered groups of EPCs. In particular, three graphical methods are integrated:

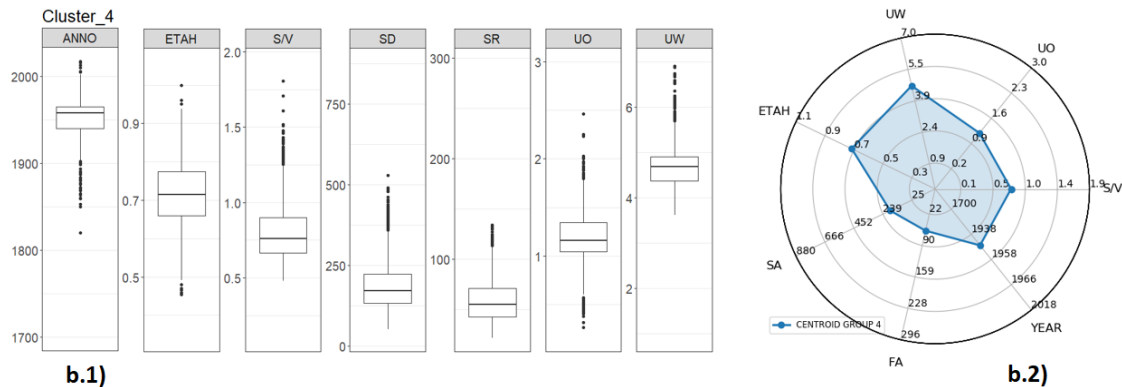
- (1) *Boxplot distribution.* A boxplot (also known as whiskers plot) is a standardized way of displaying the data distribution through their quartiles. In TUCANA, the boxplot characterizes each cluster content in terms of distribution of the values of variables considered. TUCANA includes a boxplot for each variable considered in the clustering phase, separated for each cluster, to allow the end-users to easily compare the distributions of the various attributes within each cluster.
- (2) *Radar chart.* The cluster centroids are displayed by TUCANA using the *radar chart* visualization. A radar chart is a way of comparing multiple quantitative variables. This makes them useful for seeing which variables have similar or dissimilar values, making them ideal for displaying performance. Each variable is arranged radially, with equal distances between each other. Each variable value is plotted along its individual axis and connected to the others to form a polygon.
- (3) *Decision tree.* To better explain the content of each cluster groups and to assess the goodness and robustness of the clustering results, TUCANA also trains a Classification And Regression Tree (CART) classifier [38] using as input the same features used in K-means, and using as target the labels assigned by the clustering algorithm. A decision tree is a flowchart-like structure in which each node verifies a condition on a given attribute and each branch descending from this node represents one of the possible test results, while each leaf node represents the label outcome. The paths from root to leaf represent classification rules able to easily characterize each cluster content.

Examples of boxplot, radar chart and CART are well illustrated in Figures 2 and 3.



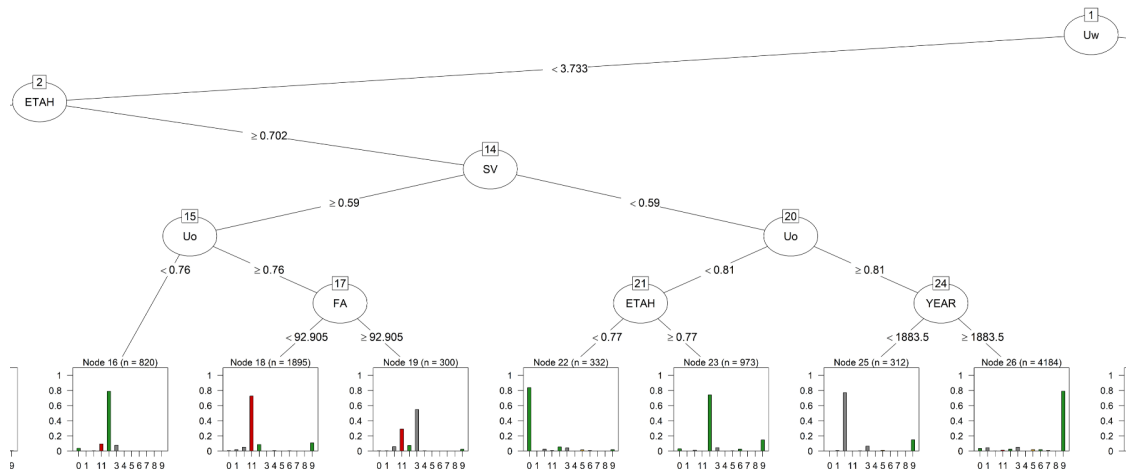
(a) Boxplot Distribution (a.1) and radar chart visualization (a.2) for Cluster 0

Figure 2. Cont.



(b) Boxplot Distribution (b.1) and radar chart visualization (b.2) for Cluster 4

**Figure 2.** Cluster 0 in (a) represents a group of high performance buildings, while cluster 4 in (b) represents a group of low performance buildings.



**Figure 3.** Example of CART generated by the framework.

### 3.2.3. Data and Knowledge Visualization

This component offers dynamic geospatial maps able to display relations between variables that influence the dwellings' energy efficiency at different spacial granularity levels. In particular, four types of geospatial maps are integrated in TUCANA:

**Choropleth maps.** Each neighborhood is colored according to the majority frequency value of the considered variable. Specifically, each variable is divided into quartiles, and then TUCANA simply counts the frequency of each quartile-bin and generates a single candidate value, which is the majority value for each neighborhood.

**Scatter maps.** Each dwelling is represented on the map as a point, which color represents the value of the selected variable.

#### Choropleth and scatter maps.

These maps show the previous two representations together. In this way, the analyst is able to jointly analyze the relation between two variables in the same neighborhood.

**Cluster-marker maps.** Here, the features gathered by the the EPCs belonging to a specific area are aggregated according to the majority value of the cluster results. These maps are able to provide an aggregate view of the EPCs and are the multidimensional result of the analysis carried out on the attribute selected. Dwellings with similar characteristics are grouped in a circle at a less detailed zoom level. In this way, TUCANA is able to represent in a 2D map, with the high-dimensional results returned by the multivariate clustering phase.

### 3.2.4. Web Application

To make a decision that takes all available knowledge into consideration, it is necessary to bring the stakeholders together and to make them share their knowledge. To this aim, in TUCANA, a web application has been integrated to share the extracted knowledge to the different stakeholders, integrating innovative and attractive functionality that aims at helping the potential users in their decision-making process.

### 3.3. Predictive Modeling

Generalization usually refers to a machine learning model's ability to perform well on new unseen data rather than just the data that it was trained on. Since the available data may not include the entire area considered for analysis, an automatic process of generalization of the extracted knowledge items could help the analyst to generalize the main relevant insights. In TUCANA, two types of generalization have been implemented: (1) *fine-grain*, to predict the value of a specific variable of interest that is missing or inconsistent because wrongly inputted by the analyst within the EPC of a building (this is a rare event affecting less than 1% of the EPCs analyzed); and (2) *coarse-grain*, to predict the features of a building for which no EPC is present, based on the features assumed by similar buildings.

#### **Fine-grained generalization.**

This generalization process includes two steps which are sequentially done: (1) a prediction phase; and (2) a classification phase. During the prediction phase, a regression machine learning algorithm is applied to predict the value of one missing cluster input variable. Specifically, TUCANA performs the following steps:

1. A regression model is built on the cleaned dataset by analyzing a subset of cluster input variables.
2. A grid-search algorithm is applied to find the best regression algorithm (i.e., linear regression [39], LASSO regression [40], polynomial regression [41], and k-Nearest Neighbors Regressor [42]) and the most suitable input parameters to gather the best performance through a K-fold validation.

We defined this methodology as the *fine-grained* approach, as it allows the prediction of some missing or inconsistent values variables within a new EPC. This strategy improves the data cleaning step, as missing or inconsistent record related to a specific variable in some certificates can be estimated and then included or replaced, avoiding to discard the entire EPC from subsequent analysis. In this way, the number of available EPCs could increase significantly, allowing to generate a fine-grained energy mapping even for those cities where the quality of available data is not high. TUCANA integrates this procedure to retrieve EPCs that present a missing or an out of the admissibility range input variable as defined in the cleaning phase.

**Coarse-grained generalization.** The classification phase is then integrated to summarize the energy performance of the new dwelling without performing the entire analysis. TUCANA automatically predicts for a new dwelling the cluster label, representing its performance through a spatial constrained k-Nearest Neighbors, an extension of the K-NN algorithm proposed in Cover et al. [4]. The novel aspect is related to a solid constraint enforced to locally define a specific neighborhood for each building under analysis. Precisely, this novel algorithm consists of the following two steps:

1. Identification of the **dwelling neighborhood** given a threshold (i.e., maximum number of dwellings). Given the latitude and longitude of the new dwelling, its closest dwellings (in terms of Euclidean distance) are selected to define its neighborhood (Step 1).
2. **K-nearest neighborhood phase.** Among the selected neighbors, the top K similar EPCs (according to the available cluster input variable) are chosen (Step 2A). Among those selected in Step 2A, the cluster label to be predicted is given by the most frequent label according to a majority voting model (Step 2B).

The above methodology is exploited in TUCANA when:

- All EPC features (considered in the cluster analysis) are available for the new dwelling.
- Only the geometrical features (i.e., a subset of features considered in the cluster analysis) are available for the new dwelling.
- Only latitude and longitude are available for the new dwelling. Specifically, in this particular case, TUCANA defines the dwelling neighborhood (Step 1) and then computes the majority model to predict the cluster label (Step 2B). Only Steps 1 and 2B are carried out.

In this case, the strategy has been defined as coarse-grained as it only allows the prediction of the cluster label to which a new building belongs. This strategy allows to know in real time the label of a new building (e.g., building whose EPC has just been released), without having to redo the whole clustering analysis. In this way, the energy mapping of the city could increase significantly, improving the impact on the decisions of the different stakeholders involved.

#### 4. Experimental Results

The purpose of this section is to show the experimental results obtained by applying TUCANA on a real use case (see Section 4.1). The analyses performed and the corresponding discussion are reported in different sections. Section 4.2 shows how the real data were pre-processed to be suitable for later analysis. Section 4.3 shows the results of the clustering phase, grouping together EPCs with similar characteristics. Section 4.4 shows the results obtained through the use of different interactive maps. Finally, Section 4.5 discusses how well the fine-grained and coarse-grained approaches perform on the real data under analysis.

TUCANA was developed in Python [43], including the scikit-learn library (for the analytic steps) and the folium library (to implement the visualization phase based on maps). The web application was built in Flask [44], which is a micro web framework written in Python.

##### 4.1. The Case Study Description

TUCANA was experimentally tested and validated on a real collection of EPCs generated in the years between 2009 and 2018 in the Piedmont Italian region. The dataset has been gathered and openly released by CSI Piemonte (the regional Information System Consortium) (<http://www.csipiemonte.it/web/it/>) [45], and regulated by the Piedmont Region authority (Sustainable Energy Development Sector). The dataset includes nearly 270,000 EPCs.

Basic statistic distributions related to the entire dataset are reported in Figure 4. Specifically, Figure 4a shows the frequency distribution of the EPCs for each year of issue through the pie chart representation. The distribution is skewed and the most frequent years are 2010, 2013, and 2017. The legend is sorted starting from the top of the pie chart plot. In Figure 4b, the frequency of the EPCs for each province is reported. Turin city includes more than 50% of EPCs of the original database. The intended use percentage of the dataset is reported in Table 1, adopting the classification described in the D.p.r. n. 412/1993 [46]. Due to this basic statistics, here we discuss how well TUCANA performs on EPCs related to residential building (i.e., Intended use = E1(1)) of the city of Turin, collected from all years of issue. The Turin dataset includes 47,623 EPCs. With the domain expert, a subset of available features (S/V, Uo, Uw, SA, FA, Year, and ETAH) were selected for the next analytics steps, as reported in Table 2. The experiments were performed exclusively on attributes that can be categorized as geometric, thermo-physical, and system based features. The geometric and thermo-physical variables are real life variables collected through surveys and inspection from energy experts. The system based variable (i.e., ETAH) is evaluated on the basis of precalculated values of efficiency of each subsystem considering the real generation system, distribution network, terminal unit, and control system installed in each building. In this table are also reported, for each variable, the admissibility ranges defined by the domain expert.

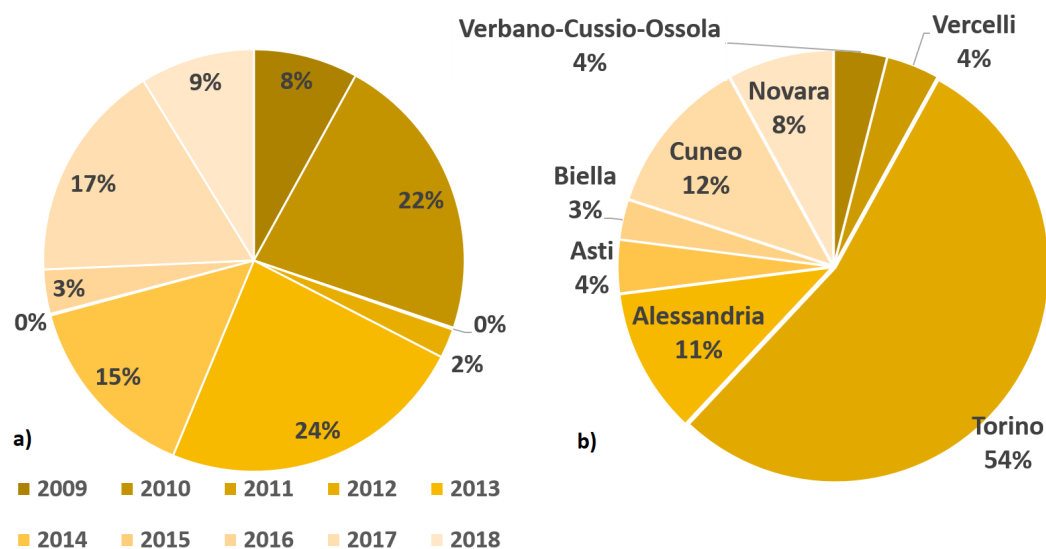
**Table 1.** Intended use percentage distribution.

Intended use	E1 (1)	E1(2)	E1(3)	E2	E3	E4 (1)	E4 (2)	E4 (3)	E5	E6 (1)	E6 (2)	E6 (3)	E7	E8
Percentage %	84.93	3.18	0.05	3.18	0.12	0.07	0.04	0.84	5.02	0.02	0.09	0.04	0.19	2.21

**Table 2.** Selected attributes characterizing EPCs and corresponding validity ranges.

Names	Acronym	Units	Min	Max
Aspect ratio	S/V	[m <sup>-1</sup> ]	0.1	2
Surface area	SA	[m <sup>2</sup> ]	24.9	880
Floor area	FA	[m <sup>2</sup> ]	21.5	296
Average U-value of the vertical opaque envelope	UO	[W/m <sup>2</sup> K]	0.15	3
Average U-value of the windows	UW	[W/m <sup>2</sup> K]	0.9	7
Heating system global efficiency	ETAH <sup>1</sup>	-	0.3	1.06
Construction year	Year	-	1700	2018

<sup>1</sup> This index takes into account the efficiency of each subsystem of the building, including generation, distribution, and control. The buildings equipped with heat pumps as heating generation system were preliminary excluded considering that these systems are installed for a very limited number of residential buildings in the available dataset. In fact, the large majority of residential buildings in Italy use gas boilers as heating generation system.

**Figure 4.** Basic statistic distributions considering before data pre-processing.

Numerous experiments were carried out to evaluate the effectiveness and robustness of TUCANA in: (1) correctly separating groups of buildings characterized by similar features; (2) visualizing the results through interactive maps with different granularity levels, making it usable also to different stakeholders due to the web application; and (3) correctly inferring missing or inconsistent features of new dwellings through a generalization process.

In the data-preprocessing phase, the univariate outlier detection through gESD was performed setting the number of outliers to 0.5% of the certificates and the critical value to 0.01; the range of values on which the DBSCAN minPoints parameter was tested was set to [2, 20]. In the clustering phase, to find the best configuration of the k-means, all k values in the range [2, 50] were tested. Finally, the best parameter configuration for the spatial-constrained K-NN was found by setting the maximum number of buildings to be considered as neighbors to 2000 and the maximum value that k can assume to 1000, and testing the various combinations of solutions, while the parameters for the regression algorithms were configured thanks to a grid-search technique, considering the possible k values for

the K-NN regressor in a range [1, 10] and the following as possible  $\lambda$  values for the lasso regressor: [0.00001, 0.0001, 0.001, 0.01, 0.1].

#### 4.2. Data Pre-Processing

The TUCANA preprocessing phase includes: (i) *geospatial data cleaning*; (ii) *outlier detection and removal*; (iii) *feature correlation and normalization*; and (iv) *definition of the acceptability range* to prepare the real data to be effectively analyzed through the subsequent analytics steps.

**Geospatial data cleaning.** To clean the geospatial coordinates included in each EPC, TUCANA applies the algorithm proposed in Section 3.1. Specifically, TUCANA automatically corrects all the address information (i.e., address, house number, ZIP Code, latitude, and longitude) with inconsistency in their values. To verify the correctness of the data, the values in the EPCs were compared with the values in an open dataset [47] (available at: [https://sciamlab.com/opendatahub/dataset/c\\_l219\\_260](https://sciamlab.com/opendatahub/dataset/c_l219_260)) provided by the city of Turin which contains all the addresses, ZIP codes, and geographical coordinates of the city. For the addresses for which it was not possible to correct the inconsistencies with the algorithm used, TUCANA requested a call using Google Geocoding APIs, in order to reconstruct EPCs' addresses and coordinates. Thanks to both solutions, 99% of the addresses were resolved correctly.

**Outlier detection and removal.** To achieve good results, an intense data cleaning session was performed before the knowledge extraction phase. To do this, TUCANA integrates different strategies to support the analyst into correctly identify and remove possible outliers.

As discussed in Section 3, TUCANA relies on gESD and analysis of frequent data distribution to remove outliers feature by feature, while DBSCAN allows identifying outliers through a multivariate analysis. To help the analyst configure input parameters, Figure 5 reports the analysis for automatically set the value of *MinPoints* and *Epsilon* using the k-distance plot, which shows the distances (y-axis) of each point of the dataset (x-axis) from its nearest kth point. These distances are ordered in descending order. TUCANA performed many runs and the result is that the curve was quite stable in the values  $k = 5$ . The eps value was chosen by looking at the elbow of the plot shown in Figure 5; it is important to choose an appropriate value, because, with a too large eps value, the points will end up in the same cluster, while, with a too small eps value, most of the points will be considered noise. TUCANA sets as good possible configuration for the DBSCAN algorithm  $\text{MinPoints} = 5$  and  $\text{Eps} = 0.28$ .

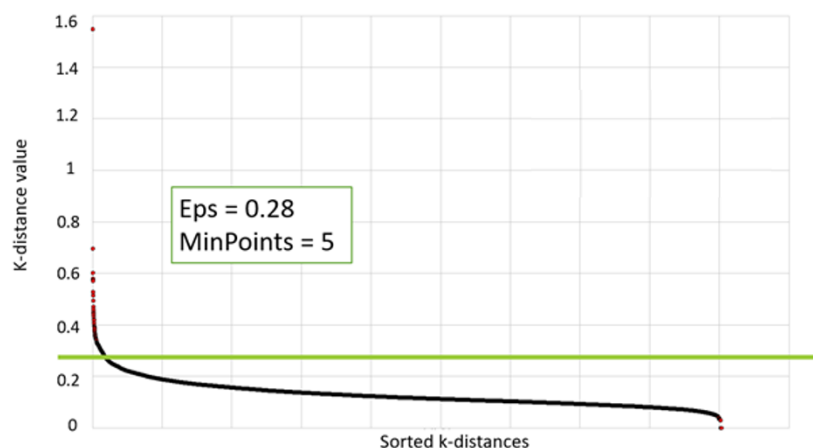


Figure 5. K-distance plot for the DBSCAN tuning parameters.

**Feature Correlation and Normalization.** Before clustering, the correlation between the considered numerical attributes is computed to discover possible correlated attributes. Figure 6 shows the correlation matrix, containing the Pearson's coefficient for each pair of variables considered. Cell colors represent the degree of correlation between two variables (the darker the cell, the higher the linear correlation). In this case, we can see that all variables are weakly correlated (the highest value is

0.62) and therefore there is no strong linear dependency between a pair of attributes. Thus, none of these was discarded and all these variables (i.e., S/V, Uo, Uw, SA, FA, Year, and ETAH), after being normalized, were used in the clustering phase to discover groups of similar buildings.

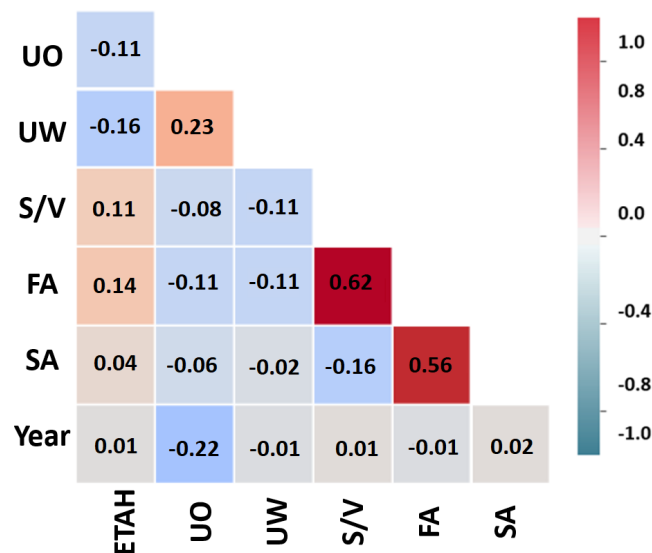
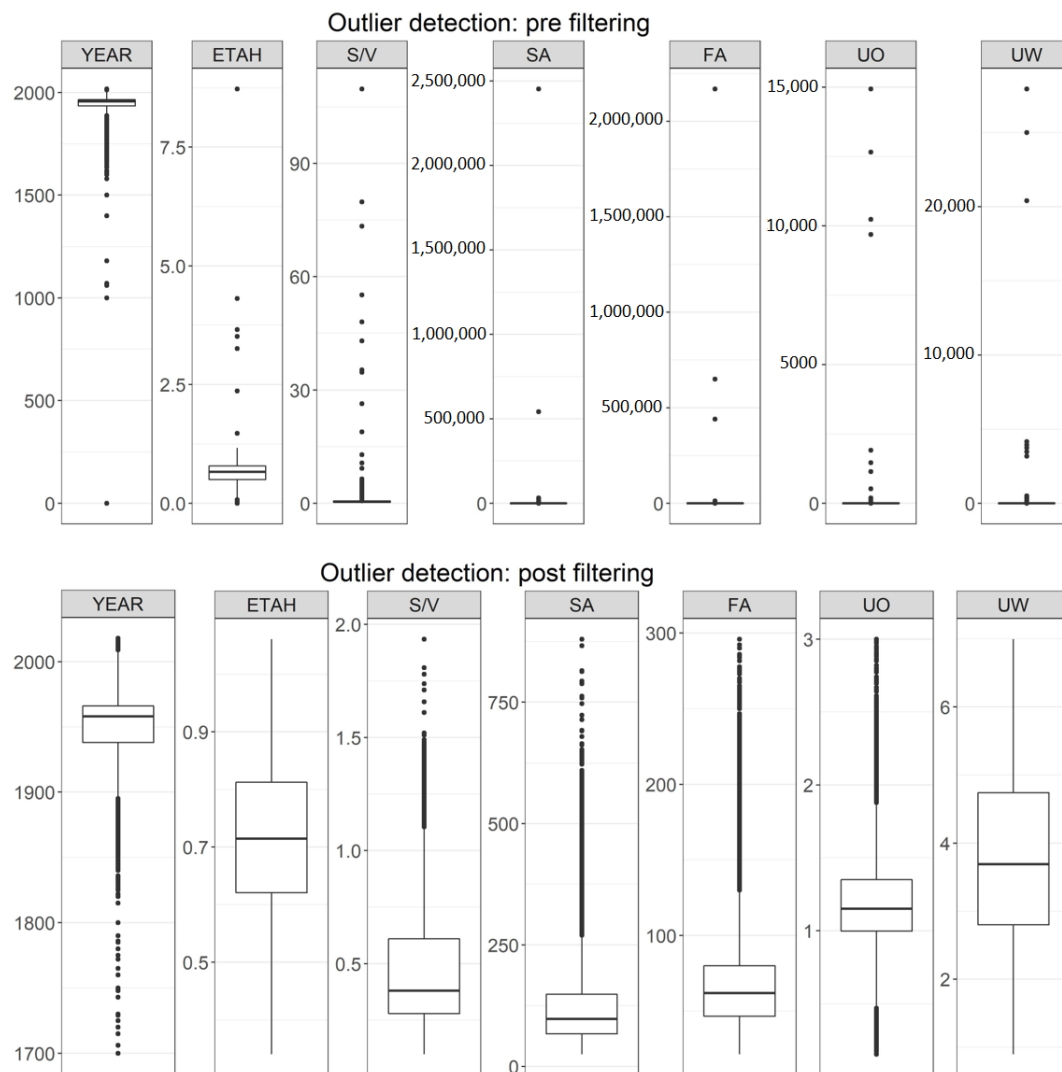


Figure 6. Correlation matrix on the selected features.

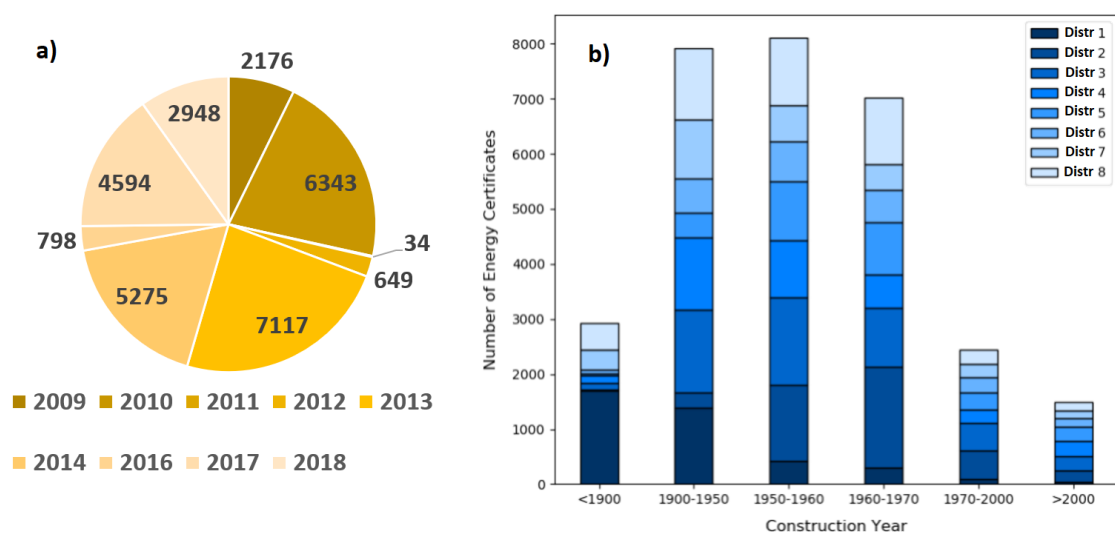
**Definition of the acceptability range.** Table 2 shows the selected attributes and their corresponding units, together with the validity ranges defined by the experts for each of these attributes. The choice of these limits is the consequence of an iterative approach based on data-driven analysis supported by domain expert advice. The obtained results (e.g., one-dimensional and bi-dimensional analyses) were shown to the experts, who, thanks to their expertise, advised us on appropriate values to use as limits. This process was repeated several times, with different analyses for each attribute, slightly modifying the various limits each time. This led to the definition of the values shown in Table 2, which represent the final result of this process.

TUCANA automatically removes EPCs characterized by at least one value outside these defined ranges: after this operation, the resulting dataset includes 29,934 EPCs. Figure 7 reports the effect of the pre-processing phase before and after the outlier detection and removal.

The dataset characterization after the pre-processing phase is reported in Figure 8. Specifically, Figure 8a reports the pie-chart distribution of the number of EPCs for each year under analysis. The pie chart shows how 2010 and 2013 were the years with the highest emissions of energy certificates, perhaps due to some energy saving incentive campaigns or similar initiatives. Figure 8b reports instead the stacked bar graph of the number of EPCs for each district of the Turin city divided for each construction year of the dwellings. The stacked bar shows that in all the districts we have buildings built in different years with different distributions. This characteristic is true and therefore it can be concluded that the sample analyzed represents a real case.



**Figure 7.** Effect of the pre-processing phase before and after the outlier detection and removal through the boxplot representation.



**Figure 8.** Basic statistic distributions after data pre-processing.

### 4.3. Descriptive Modeling

To discover groups of buildings cohesive and well separated, TUCANA relies on the K-Means algorithm. For each  $k$  value in the defined range, the SSE index is evaluated and plotted against the  $K$  value, as reported in Figure 9. TUCANA automatically selects as optimal value the one that represents the point of maximum curvature. Based on this method, TUCANA selects  $K = 12$ .

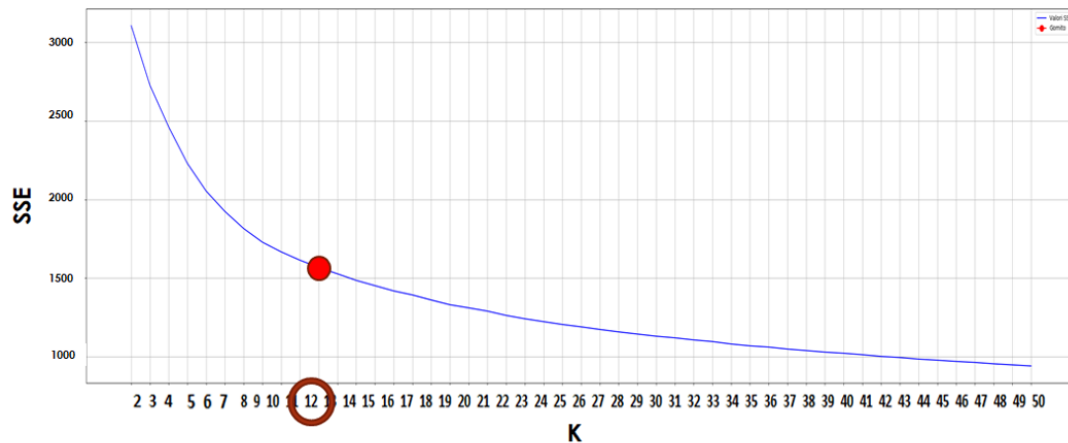


Figure 9. SSE trend against  $K$ .

Table 3 reports the cluster cardinalities, separated for each district. Clusters 8 and 9 are the largest groups, with 16.4% and 12.4% of buildings, respectively, while Clusters 3 and 10 are the smallest ones, containing 2.8% and 2.6% of buildings, including all historical buildings. Obtained results are perfectly in line with the building distribution reported in Figure 8b. Thus, the cluster analysis correctly grouped buildings based on thermo-physical properties and geometrical characteristics and system efficiencies.

Table 3. Cardinality of the cluster set, separated for each district.

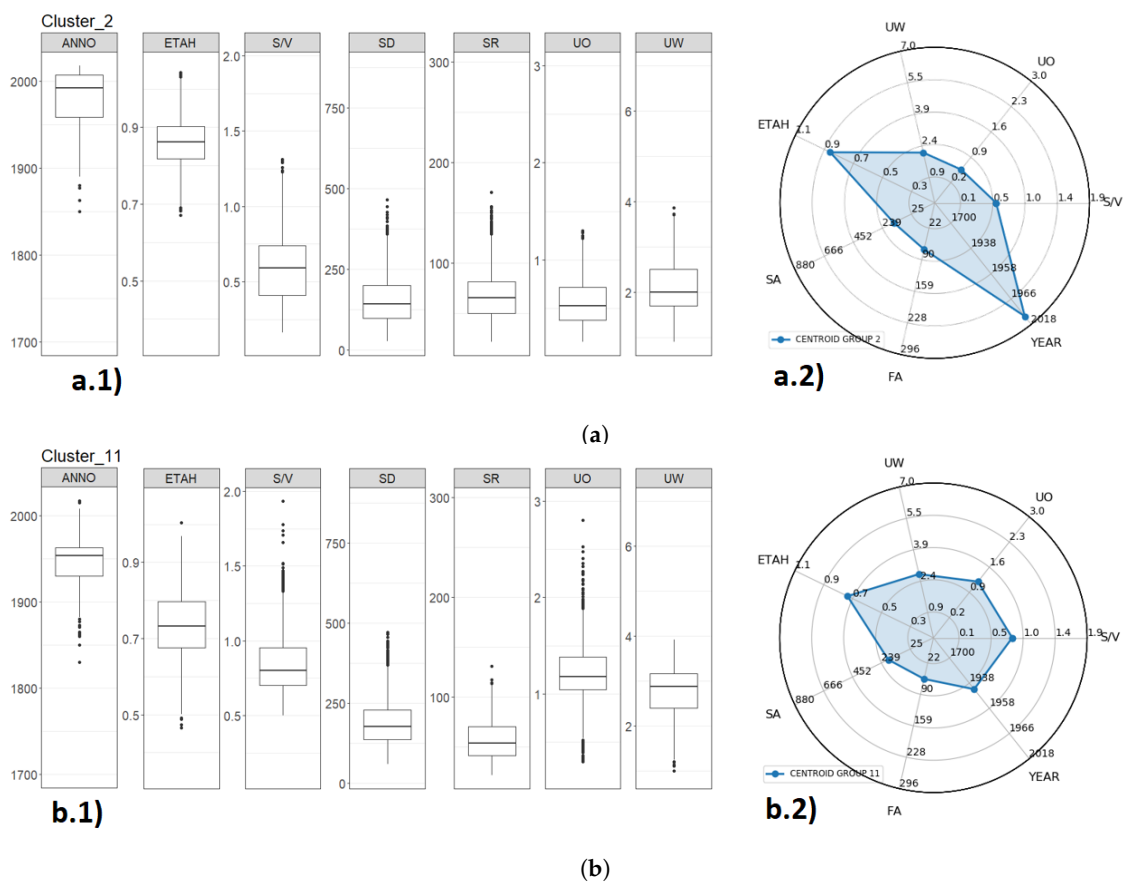
Cluster ID	Districts								Total EPC
	1	2	3	4	5	6	7	8	
Cluster 0	101	245	321	217	281	222	172	224	1783
Cluster 1	231	289	311	249	131	137	145	317	1810
Cluster 2	91	236	264	283	262	111	196	240	1683
Cluster 3	251	54	92	79	23	42	109	207	857
Cluster 4	218	395	523	304	306	270	291	413	2720
Cluster 5	430	185	234	165	33	37	105	261	1450
Cluster 6	383	758	688	472	375	297	360	750	4083
Cluster 7	419	433	637	480	415	325	351	514	3574
Cluster 8	435	738	860	649	587	450	496	701	4916
Cluster 9	480	591	643	472	351	274	359	555	3725
Cluster 10	643	2	8	14	1	9	53	78	808
Cluster 11	255	321	440	245	300	268	292	404	2525
									29,934

Each cluster is then locally characterized through boxplots and radar charts to provide some human-readable insights cluster by cluster. Specifically, here we compared boxplots and radar charts of some homogeneous groups of low and high performance buildings and we offered an example of rules obtained from the CART generation.

Figure 2 (a.1 and b.1) shows the boxplot distributions for each variable for Clusters 0 (Figure 2a) and 4 (Figure 2b). Specifically, Cluster 0 presents fairly recent dwellings with aspect ratio, surface area,

and average U-value of the windows tending to lower values, compared to the values characterizing Cluster 4. This means that the buildings belonging to Cluster 0 seem to be better performing than those belonging to the ones to Cluster 4. This trend is also confirmed by the polygons generated in the radar chart visualization in Figure 2 (a.2 and b.2). Each vertex of the each polygon represents the corresponding centroid value of each variable. This simple visualization is useful to easily compare different cluster centroids, thus the corresponding groups.

On the other hand, Figure 10 shows the comparison of other two clusters, one representing high performance buildings and the other low. Cluster 2 (see Figure 10a) has a high value of both ETAH and year of construction, and it is characterized by lower values of average U-value of the vertical opaque envelope and windows and aspect ratio. This features usually characterizes high performance buildings. Figure 10a shows instead a group of low performance buildings with different characteristics to the ones presented in Figure 10b. Indeed, it presents lower values of ETAH and year of construction and higher values for the other features. This means that Cluster 2 includes dwellings with good geometric and thermo-physical properties, while Cluster 11 takes into account older buildings characterized by a poorly performing envelope. High-performance buildings and low-performance buildings have been divided into two subgroups (Clusters 0 and 2 for high-performance and Clusters 4 and 11 for low), because each cluster is characterized by different values, as the buildings that compose it have different thermo-physical and/or geometric characteristics from the buildings in the other clusters. For example, Cluster 2 is characterized by higher values of ETAH and lower values of average U-value of the vertical opaque envelope and windows than Cluster 0 (even though they both contain high performance buildings).



**Figure 10.** Boxplot distribution (a.1 and b.1) and radar chart visualization (a.2 and b.2) for Clusters 2 and 11: (a) group of high performance buildings; and (b) group of low performance buildings.

As discussed in Section 3.2.2, each cluster is also characterized by rules extracted from CART to provide some additional insights. The decision tree presented in Figure 3 represents a portion

of the entire CART returned by TUCANA. This kind of extracted knowledge is intuitive and human-readable. CART is executed on the building partition discovered through the cluster analysis. The input parameters were tuned to reach an accuracy of 80%. The paths from the root to each leaf node represent classification rules that could be used to summarize in a compact way the features of each cluster. This methodology provides the analyst to determine some *if-then* rules characterizing the homogeneous groups that have been generated in the clustering phase. An example of extracted rule for Cluster 0 is reported in Figure 3; specifically, if  $\{UW < 3.733, ETAH \geq 0.702, SV < 0.59, UO < 0.81, ETAH < 0.77\} \Rightarrow \{\text{Cluster Label} = \text{Cluster 0}\}$ .

Based on the rich-set of human-readable insights characterizing each group of dwellings, with the support of the domain experts, we shortly described each group, as reported in Table 4. In addition, different colors are used to mark the different performance labels of each cluster: green represents dwellings with high performances, orange represents dwellings with medium performances, and red represents dwellings with low performances. These colors are then used to color the navigable maps. It is interesting to notice that the same energy label could be defined by different combinations of input variables, including thermo-physical and geometrical aspects. If it is not possible to define a specific energy performance label, the grey label Not Classified (NC) is given. For example, Cluster 10 is characterized by historical buildings and a single description is not achievable.

**Table 4.** Semi-supervised data labeling.

Cluster ID	Performance	Description
0	High	High performing envelope, medium performing energy system
1	NC	Low performing envelope, low values of SV
2	High	High performing envelope and energy system
3	NC	Buildings with large surface area
4	Low	Low performing envelope, high values of SV
5	Medium	Low performing envelope, medium performing energy system, low values of SV
6	High	Low performing envelope, high performing energy system, low values of SV
7	Medium	High performing envelope, low performing energy system, low values of SV
8	Medium	Medium performing envelope, low performing energy system, low values of SV
9	High	Medium performing envelope, medium performing system, low values of SV
10	NC	Historical buildings
11	Low	Medium performing envelope, medium performing system, high values of SV

#### 4.4. Data and Knowledge Visualization

In this subsection, the methodologies adopted and integrated in TUCANA are illustrated to present how the results obtained in the previous analytics phases are shown to the different stakeholders. In particular, two different visualization tools are proposed and developed: (i) different geo-localized data representations through the use of dynamic and easily navigable maps; and (ii) the design and development of a web application for visualizing the knowledge acquired during the analysis.

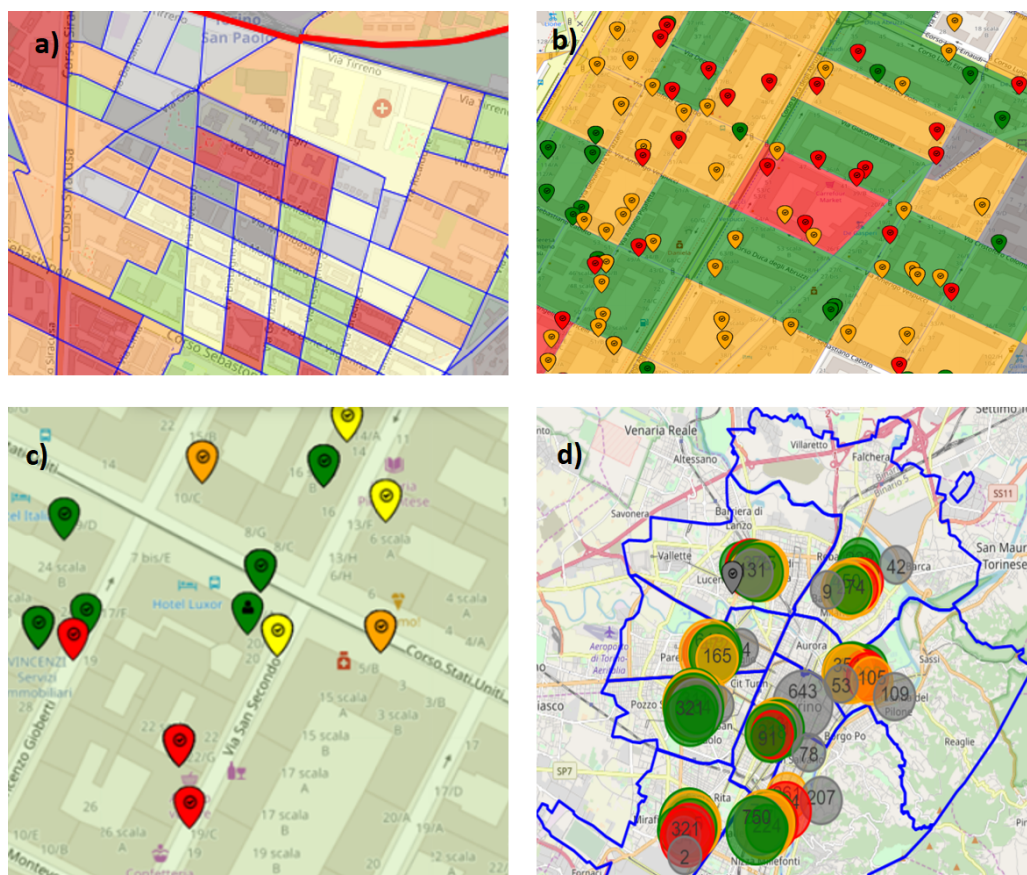
The main objectives of these two types of visualization are:

- enhancing the extracted knowledge by providing a correct interpretation of the available data, using an interactive data navigation and exploration;
- interpreting the knowledge at different levels of detail and transform data into useful knowledge in order to guide the processes of decision-making;
- reporting the different details of the extracted knowledge to different stakeholders; and
- making the EPCs data easily usable and effective to support the decision making process at any time.

##### 4.4.1. Dynamic Geospatial Maps

Since energy certifications are characterized by geographical attributes, they can be represented on dynamic and geo-located maps. This type of representation makes it possible to obtain immediate information on the efficiency of a dwelling or groups of dwellings, as well as on the energy performance of specific areas. In TUCANA, four types of different maps were integrated, as reported in Section 3.2.3.

The proposed maps visualize variable resolution details, showing the user the dwellings' performances at different spatial granularities. Specifically, Figure 11a represents the choropleth map at the neighborhood zoom level of the Floor Area input variable. In particular, the city of Turin is divided into blocks. Each block is colored according to the majority values of the selected variable. This high-resolution map allows different end-users to easily assess the energy efficiency of potentially interesting sites. This visualization allows a univariate analysis at neighborhood level. On the other hand, Figure 11c shows the smallest granularity level through the scatter map. It reports a point for each EPC (representing a residential unit) contained in the selected area. The corresponding color of each point represents the value of the cluster results after the labeling of each group. In TUCANA, it is also possible to combine them to analyze how a variable influence the cluster label, as shown in Figure 11b. Lastly, similar to the choropleth maps, the cluster-marker maps aggregate multiple EPCs coloring the dynamic markers according to the majority of the values of the aggregated points. This kind of map is new and it is proposed to visualize the areas presenting lower and higher energy performances at district or city levels. The cardinality of each cluster is reported in the cluster-markers, while the majority value of the cluster label is used to color the marker itself. The cluster-marker visualization is proposed to show the cluster analysis results on a map. To this extent, TUCANA integrates cluster-markers to introduce a new feature to the maps which represents the results of the exploratory analysis, in order to characterize the energy efficiency of several dwellings through various geometrical and thermo-physical properties. The marker is colored based on the label defined in Table 4 and models the combination of features representing each EPC. The number reported in the big circle is instead the number of buildings aggregated in the marker. Details of each marker (cluster ID, range of each feature) are shown in the legend.



**Figure 11.** Examples of: choropleth map (a); choropleth and scatter map at neighborhood level (b); scatter map (c); and marker-cluster maps at city level (d).

#### 4.4.2. Web Application

The main purpose of the web application is to make available the extracted knowledge to different types of users. In addition, both geolocalized maps and basic statistics on open EPCs data are integrated. The realization of this tool was preceded by an in-depth learning phase on how to report the information to different stakeholders. To streamline the complexity of the data analytics and allowing all people (including those who never study data science algorithms) to benefit from this analysis, different stakeholders have been integrated to try to make the most of the extracted knowledge and calibrate it for each end-user, based on the corresponding expertise. Specifically, the website could support three main types of users:

- **Private citizens:** They could be interested in knowing the building performance of some areas of the city of Turin, using and consulting the site as a guideline tool if interested in buying or renting a property.
- **Public administration:** Oriented towards the energy efficiency of dwelling structures, it could use the information presented to identify the critical areas of the city where improvements can be made, thus possible target of specific tax incentives for energy improvements.
- **Energy service providers:** They could be interested in knowing the current energy performance of the city capturing the high-level overview of heating energy demand at a city level or district, and deepen the knowledge of the specific building. This kind of knowledge could help the definition of specific marketing campaigns to promote a subset of services only to interested users.

In Figures 12–15, different screen-shots of the web application are reported. Specifically, different statistics are reported in Figure 12 to help the end-user analyze the distribution of specific features of analysis. The reported statistics are different for each stakeholders, since the level of expertise could be different. The user can navigate the map and after the selection of the area of interest, the statistics are computed and graphically shown. Figure 14 reports an example of choropleth map related to the surface area feature. The different stakeholders could analyze at different level of detail (building, neighborhood, district, and city) the area of Turin. Scatter and marker-cluster maps are reported in Figures 14 and 15, respectively. In this way, end-users can prepare for a more focused and aware navigation of the components present in the web application.

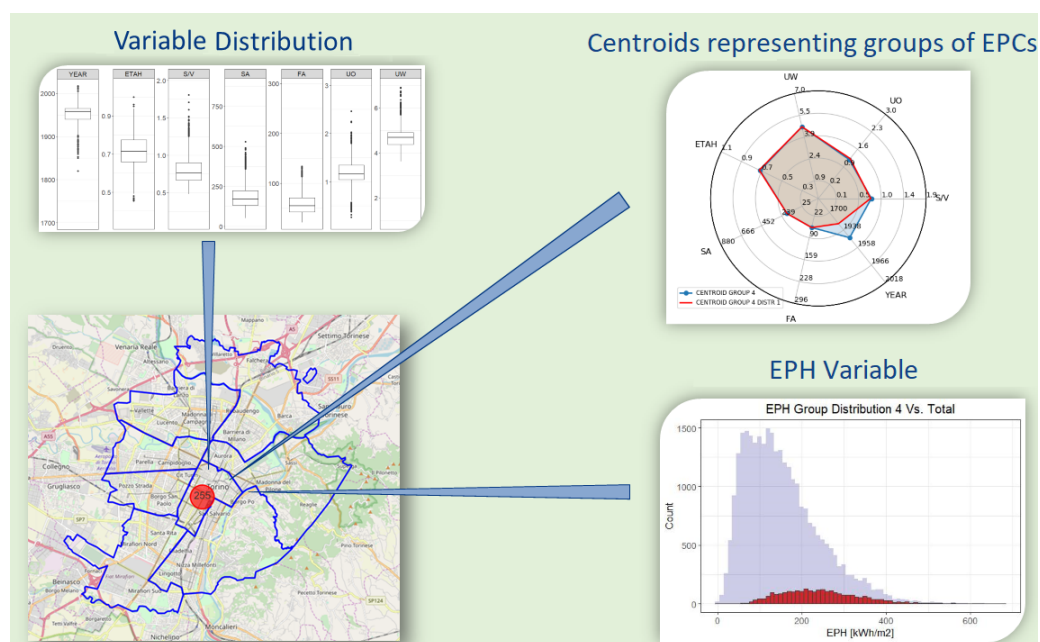


Figure 12. Example of statistics for a cluster.

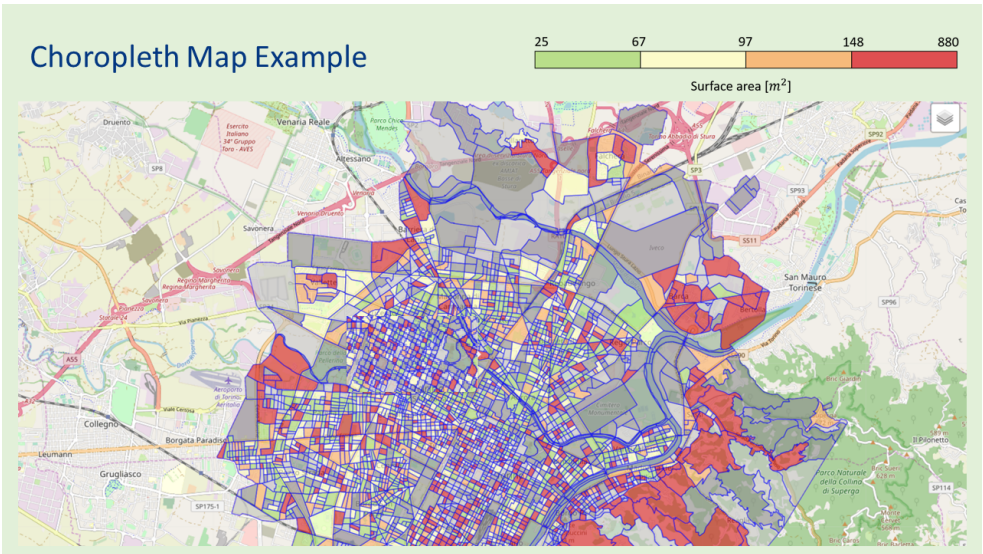


Figure 13. Example of choropleth map for the surface area feature.

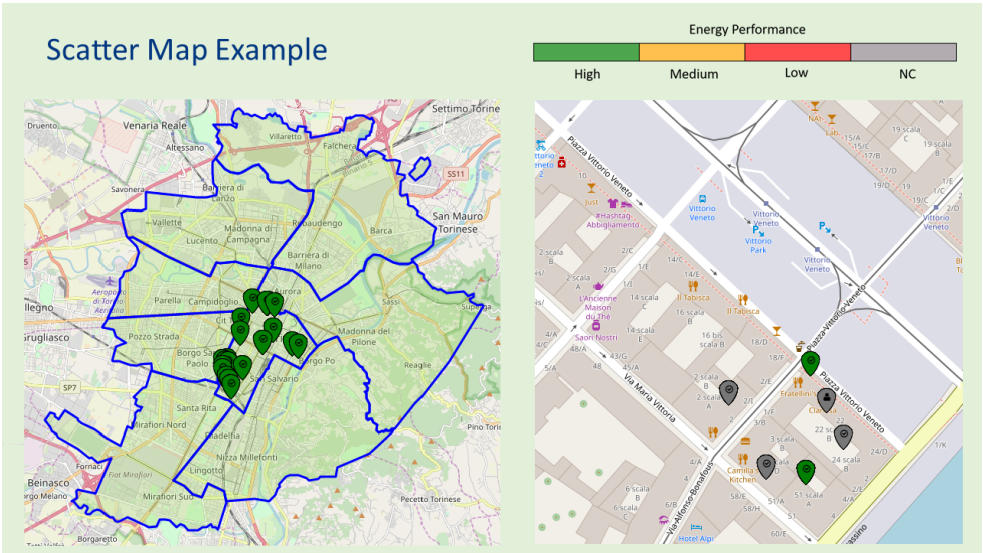


Figure 14. Example of scatter maps at: city level (left); and district level (right).

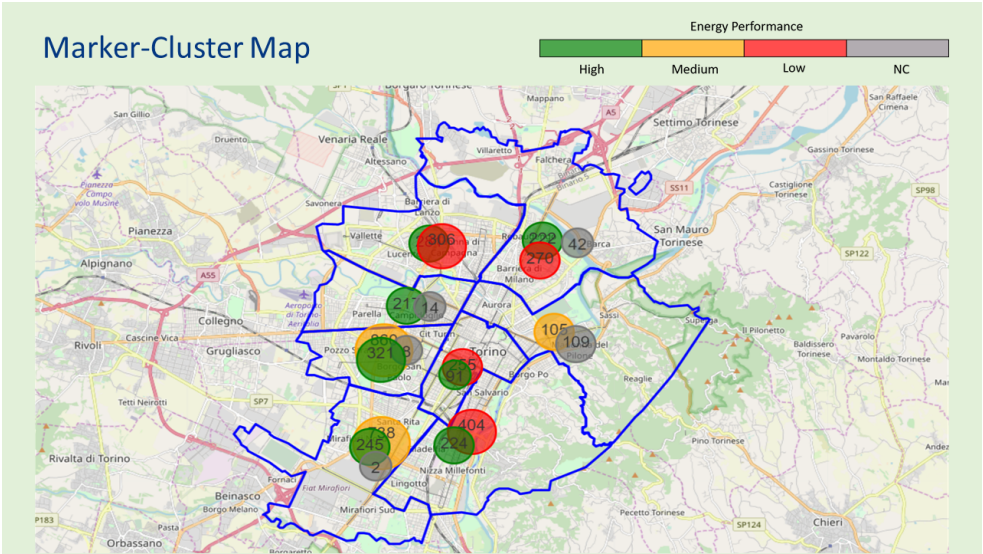


Figure 15. Example of marker-cluster map at city level.

#### 4.5. Predictive Modeling

TUCANA proposes a methodology to generalize the extracted knowledge from EPCs with the aim of predicting the energy behaviour of dwellings that are not present in the available database. To achieve this goal, two possible predictions are managed: (1) *Coarse-grained*: The case in which the prediction of the cluster label for a new building is shown. In particular, it is assumed that the new certificate has all the attributes of interest or only the three geometric variables (i.e., S/V, FA, and SA). Based on the available data, a classification task is performed. (2) *Fine-grained*: The case in which the new energy performance certificate does not have one of the seven variables used for the characterization of the extracted knowledge. Thus, through a regression task, missing or inconsistent values are estimated.

##### 4.5.1. Coarse-Grained

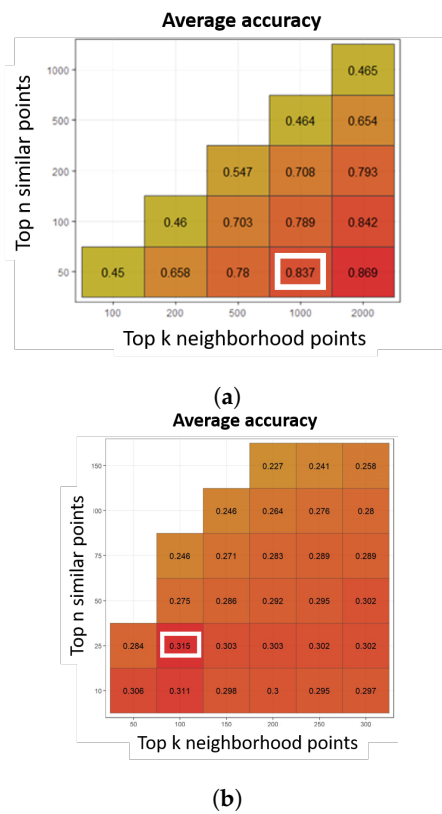
The prediction of the cluster label of a new building is made in two different cases: (1) when the new EPC contains all seven variables shown in Table 2; (2) when the new EPC contains only the three geometric variables (S/V, FA, and SA) (because other non-geometrical variables may be missing or inconsistent). In both cases, as expressed in Section 3.3, the following two-level methodology is applied:

- **identification of the dwelling neighborhood**, in order to select the closest dwellings; and
- **K-NN similarity**, in order to extract the K-similar EPCs among the selected neighbors.

The prediction of the cluster label for test observation is achieved by taking the majority label within the neighborhood selected by K-NN. The default approach relies on two input parameters: the maximum number of dwellings to be considered as geographical neighborhood and the K parameter of the K-NN algorithm. To set both parameters, the so-called grid-search is implemented, which takes several possible values for the two parameters and measures the accuracy (number of correct predictions, divided by the total number of predictions) of the resulting K-NN model. Figure 16 shows all the accuracy values obtained by analyzing the different combinations of input parameters. Figure 16a refers to the case where all seven variables are available, while Figure 16b refers to the case where only the three geometric variables are present in the new EPC.

As shown in Figure 16a, we chose 1000 training observations as geographical neighborhood and  $K = 50$ , in order to have an accuracy value of 0.837. This result indicates that the model with these parameters has a good performance, and, even if the highest accuracy value is associated with a neighborhood of 2000 points, the gain is minimal compared to the chosen value. In Figure 16b, instead, there are in general values much lower than the previous case, which means that this procedure is less accurate. This reduction in accuracy is explained by the lack of thermo-physical and system based variables (UO, UW, ETAH, and Year), without which it is very difficult to predict the cluster label. In this last situation, we chose a number of geographical training neighbors equal to 100 and  $K = 25$ , because these parameters guarantee us the highest accuracy value (0.315). As expected, when only geometric variables are available, the performance of the classifier is very low.

The different performance in the two cases is also highlighted in Tables 5 and 6. These tables show the precision and recall values for each cluster label, as well as the average precision and average recall values, where *precision* is defined as  $TP/(TP + FP)$ , while *recall* is defined as  $TP/(TP + FN)$ . The first table represents the case in which all the variables were considered, while the second took into account only the three geometric variables.



**Figure 16.** Grid-search for K-NN algorithm: (a) accuracy considering all the variables; and (b) accuracy considering only geometric variables.

**Table 5.** Metrics with all the variables.

Cluster ID	Precision	Recall
0	0.917	0.576
1	0.951	0.48
2	0.942	0.662
3	0.962	0.481
4	0.897	0.861
5	0.82	0.58
6	0.839	0.95
7	0.787	0.95
8	0.842	0.989
9	0.765	0.954
10	0.963	0.792
11	0.829	0.863
<b>Accuracy</b>	<b>Average Precision</b>	<b>Average Recall</b>
0.839	0.876	0.761

**Table 6.** Metrics with only geometric variables.

Cluster ID	Precision	Recall
0	0.292	0.204
1	0.117	0.051
2	0.319	0.261
3	0.790	0.299
4	0.358	0.415
5	0.497	0.467
6	0.245	0.306
7	0.247	0.185
8	0.274	0.413
9	0.203	0.142
10	0.381	0.564
11	0.397	0.429
Accuracy	Average Precision	Average Recall
0.299	0.343	0.311

#### 4.5.2. Fine-Grained

In this case, four different regression algorithms (linear regression [39], LASSO regression [40], polynomial regression [41], and k-Nearest Neighbors Regressor [42]) were tested to estimate a missing or inconsistent value for an input attribute within the EOPC of a dwelling. As an example, let us two different cases:

*Case 1:* We assumed that an EPC is missing the variable *Aspect ratio* (S/V).

The input variables to be considered to generate the regression models are all the analysis variables summarized in Table 2, with the exception of the variable that was chosen as the response variable (S/V).

*Case 2:* We assumed the EPC is missing the variable *Heating system global efficiency* (ETAH). This variable, based on the energy expertise, can be approximated with the product of the yields of the subsystems that characterize heating systems (generation (ETAG), emission (ETAE), regulation (ETAR), and distribution (ETAD)). All four components are present in the EPC, and, for this reason, in the estimation model, we assumed that only one of these components was missing, while the other three were known. Thus, in addition to the usual input variables to be considered in this regression model, the product of the yields present in the test observation must also be added as an additional regressor. If, for example, the missing variable is the ETAE attribute, the following regression model is generated:

$$ETAH \sim S/V + UW + UO + SR + SD + YEAR + ETAG \times ETAR \times ETAD$$

To compare the performance of the various regression methods, a grid-search technique was first used, which allowed us to identify the best values of the K parameter for K-NN and the tuning parameter  $\lambda$  in the lasso regression. Then, the accuracies of the various algorithms are compared and the most performing one is chosen. In all these experiments, good  $R^2$  values are obtained, greater or equal to 0.85, where  $R^2$  represents an indicator which, starting from the regression line, synthesizes in a single value how much the analyzed quantity deviates on average from this line.

## 5. Conclusions and Future Work

The paper presents a methodology capable of analyzing and extracting useful insights from EPCs. The proposed methodology was tested on a real dataset of energy certificates collected in Turin. After a preprocessing phase, TUCANA can support the analyst in the analysis of a large collections of EPCs, by:

- identifying cohesive clusters of EPCs with similar properties;
- characterizing each cluster found with an appropriate label indicating its energy performance;
- visualizing the results in a easily and understandable way, thanks to innovative graphic techniques at different granularity levels, making the methodology usable also to different stakeholders due to the web application; and
- predicting input features of dwellings that are not present or are inconsistent in the starting database, obtaining acceptable accuracy values using classification and regression techniques.

The proposed methodology can concretely and easily help domain experts have a global view of energy efficiency by exploiting open data. The proposed maps could easily highlight the city's neighborhoods that should need improvements because of the poor efficiency buildings. To this purpose, data-driven models combined with dynamic geospatial maps are useful for quickly supporting the decision-making process of authority planners. The latter might be interested in the knowledge provided by TUCANA to plan future financial investment policies that leverage specific building features and help to devise more targeted actions to improve energy efficiency across the city. Designers could also help define the subset of buildings and where they are located, requiring specific retrofitting strategies.

However, there is still way to improve this research activity. As future works, we aim to: (1) better validate the section of predictive modeling, perhaps with the help of an expert who generates an energy certificate for a new building and then comparing that the energy label is equal to that deduced from the methodology; (2) adapt the methodology to other data domains (air pollution and weather conditions) with the aim of improving living conditions in cities; and (3) let several users test the web application, collecting their feedback and opinions and properly changing the application according to these, thus making it more user friendly.

**Author Contributions:** Conceptualization, T.C.; Data curation, T.C. and A.C.; Funding acquisition, T.C., E.B. and M. M.; Investigation, E.D.C., S.P. and D.M.; Methodology, T.C., E.D.C. and S.P.; Software, S.P. and P.B.; Supervision, T.C., A.C., E.B., M.M., S.C. and M.T.; Validation, T.C., P.B. and A.C.; Visualization, E.D.C., S.P. and D.M.; Writing—original draft, E.D.C., S.P., P.B. and D.M.; Writing—review & editing, T.C., A.C., E.B., M.M., S.C. and M.T., All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by Edison S.p.A.

**Acknowledgments:** This work was developed in collaboration with the SmartData@Polito center for Data Science and Big Data technologies, Politecnico di Torino, Italy. The authors express their gratitude to Settore Sviluppo Energetico Sostenibile Regione Piemonte and to CSI Piemonte.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References and Note

1. Guan, X.; Xu, Z.; Jia, Q.S. Energy-efficient buildings facilitated by microgrid. *IEEE Trans. Smart Grid* **2010**, *1*, 243–252. [[CrossRef](#)]
2. Recast, E. Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings (recast). *Off. J. Eur. Union* **2010**, *18*, 2010.
3. Koo, C.; Park, S.; Hong, T.; Park, H.S. An estimation model for the heating and cooling demand of a residential building with a different envelope design using the finite element method. *Appl. Energy* **2014**, *115*, 205–215. [[CrossRef](#)]
4. Cover, T.M.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [[CrossRef](#)]

5. Kaden, R.; Kolbe, T.H. City-wide total energy demand estimation of buildings using semantic 3D city models and statistical data. In Proceedings of the 8th International 3D GeoInfo Conference, Istanbul, Turkey, 27–29 November 2013.
6. Di Corso, E.; Cerquitelli, T.; Piscitelli, M.S.; Capozzoli, A. Exploring Energy Certificates of Buildings through Unsupervised Data Mining Techniques. In Proceedings of the IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, UK, 21–23 June 2017; pp. 991–998.
7. Agugiaro, G. Energy planning tools and CityGML-based 3D virtual city models: Experiences from Trento (Italy). *Appl. Geomat.* **2016**, *8*, 41–56. [\[CrossRef\]](#)
8. Howard, B.; Parshall, L.; Thompson, J.; Hammer, S.; Dickinson, J.; Modi, V. Spatial distribution of urban building energy consumption by end use. *Energy Build.* **2012**, *45*, 141–151. [\[CrossRef\]](#)
9. Ballarini, I.; Corgnati, S.P.; Corrado, V.; Talà, N. Definition of building typologies for energy investigations on residential sector by TABULA IEE-project: Application to Italian case studies. In Proceedings of the 12th RoomVent Conference, Trondheim, Norway, 19–22 June 2011; pp. 19–22.
10. Beloglazov, A.; Abawajy, J.; Buyya, R. Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. *Future Gener. Comput. Syst.* **2012**, *28*, 755–768. [\[CrossRef\]](#)
11. Amasyali, K.; El-Gohary, N.M. A review of data-driven building energy consumption prediction studies. *Renew. Sustain. Energy Rev.* **2018**, *81*, 1192–1205. [\[CrossRef\]](#)
12. Zhao, H.X.; Magoulès, F. A review on the prediction of building energy consumption. *Renew. Sustain. Energy Rev.* **2012**, *16*, 3586–3592. [\[CrossRef\]](#)
13. Luo, H.; Cai, H.; Yu, H.; Sun, Y.; Bi, Z.; Jiang, L. A short-term energy prediction system based on edge computing for smart city. *Future Gener. Comput. Syst.* **2019**, *101*, 444–457. [\[CrossRef\]](#)
14. Kalogirou, S.A.; Bojic, M. Artificial neural networks for the prediction of the energy consumption of a passive solar building. *Energy* **2000**, *25*, 479–491. [\[CrossRef\]](#)
15. Dong, B.; Cao, C.; Lee, S.E. Applying support vector machines to predict building energy consumption in tropical region. *Energy Build.* **2005**, *37*, 545–553. [\[CrossRef\]](#)
16. Berl, A.; de Meer, H. An energy consumption model for virtualized office environments. *Future Gener. Comput. Syst.* **2011**, *27*, 1047–1055. [\[CrossRef\]](#)
17. Strzalka, A.; Bogdahn, J.; Coors, V.; Eicker, U. 3D City modeling for urban scale heating energy demand forecasting. *HVAC R Res.* **2011**, *17*, 526–539.
18. Heiple, S.; Sailor, D.J. Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. *Energy Build.* **2008**, *40*, 1426–1436. [\[CrossRef\]](#)
19. Sicilia, Á.; Madrazo, L.; Massetti, M.; Plazas, F.L.; Ortet, E. An energy information system for retrofitting smart urban areas. *Energy Procedia* **2017**, *136*, 85–90. [\[CrossRef\]](#)
20. UCL Energy Institute. London Building Stock Model. Available online: <https://www.ucl.ac.uk/bartlett/energy/research-projects/2020/nov/london-building-stock-model> (accessed on 6 December 2020).
21. Evans, S.; Liddiard, R.; Steadman, P. 3DStock: A new kind of three-dimensional model of the building stock of England and Wales, for use in energy analysis. *Environ. Plan. B Plan. Des.* **2016**, *44*. [\[CrossRef\]](#)
22. Dall'O', G.; Sarto, L.; Sanna, N.; Tonetti, V.; Ventura, M. On the use of an energy certification database to create indicators for energy planning purposes: Application in northern Italy. *Energy Policy* **2015**, *85*, 207–217. [\[CrossRef\]](#)
23. Pasichnyi, O.; Wallin, J.; Levihn, F.; Shahrokni, H.; Kordas, O. Energy performance certificates—New opportunities for data-enabled urban energy policy instruments? *Energy Policy* **2019**, *127*, 486–499. [\[CrossRef\]](#)
24. Hjortling, C.; Björk, F.; Berg, M.; af Klintberg, T. Energy mapping of existing building stock in Sweden—Analysis of data from Energy Performance Certificates. *Energy Build.* **2017**, *153*, 341–355. [\[CrossRef\]](#)
25. Xiao, H.; Wei, Q.; Jiang, Y. The reality and statistical distribution of energy consumption in office buildings in China. *Energy Build.* **2012**, *50*, 259–265. [\[CrossRef\]](#)
26. Dascalaki, E.G.; Kontoyiannidis, S.; Balaras, C.A.; Droutsa, K.G. Energy certification of Hellenic buildings: First findings. *Energy Build.* **2013**, *65*, 429–437. [\[CrossRef\]](#)

27. Gangolells, M.; Casals, M.; Forcada, N.; Macarulla, M.; Cuerva, E. Energy mapping of existing building stock in Spain. *J. Clean. Prod.* **2016**, *112*, 3895–3904. [CrossRef]
28. Caputo, P.; Costa, G.; Ferrari, S. A supporting method for defining energy strategies in the building sector at urban scale. *Energy Policy* **2013**, *55*, 261–270. [CrossRef]
29. Cerquitelli, T.; Corso, E.D.; Proto, S.; Capozzoli, A.; Bellotti, F.; Cassese, M.G.; Baralis, E.; Mellia, M.; Casagrande, S.; Tamburini, M. Exploring energy performance certificates through visualization. In Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference, EDBT/ICDT 2019, Lisbon, Portugal, 26 March 2019.
30. Cerquitelli, T.; Di Corso, E.; Proto, S.; Capozzoli, A.; Mazzarelli, D.; Nasso, A.; Baralis, E.; Mellia, M.; Casagrande, S.; Tamburini, M. Visualising high-resolution energy maps through the exploratory analysis of energy performance certificates. In Proceedings of the 2019 International Conference on Smart Energy Systems and Technologies (SEST), Porto, Portugal, 9–11 September 2019; pp. 1–6. [CrossRef]
31. Juang, B.H.; Rabiner, L. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 1639–1641. [CrossRef]
32. Google. Google Maps Platform, Documentation, 2020. Last updated: 2020-11-19.
33. Seem, J.E. Using intelligent data analysis to detect abnormal energy consumption in buildings. *Energy Build.* **2007**, *39*, 52–58. [CrossRef]
34. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the KDD, Portland, OR, USA, 2–4 August 1996.
35. Ankerst, M.; Breunig, M.M.; Kriegel, H.P.; Sander, J. OPTICS: Ordering points to identify the clustering structure. In Proceedings of the ACM Sigmod Record, Philadelphia, PA, USA, 1–3 June 1999; Volume 28, pp. 49–60.
36. Tan, P.N. *Introduction to Data Mining*; Pearson Education India (Delhi): Delhi, NCR, India, 2007.
37. Satopaa, V.; Albrecht, J.; Irwin, D.; Raghavan, B. Finding a “kneedle” in a haystack: Detecting knee points in system behavior. In Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops, Minneapolis, MN, USA, 20–24 June 2011; pp. 166–171.
38. Breiman, L. *Classification and Regression Trees*; Routledge: Oxfordshire, UK, 2017.
39. Seber, G.A.; Lee, A.J. *Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 329.
40. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
41. Kleinbaum, D.G.; Kupper, L.L.; Muller, K.E.; Nizam, A. *Applied Regression Analysis and Other Multivariable Methods*; Duxbury Press: Belmont, CA, USA, 1988; Volume 601.
42. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185. [CrossRef]
43. Rossum, G. *Python Reference Manual*; Technical Report; Centre for Mathematics and Computer Science: Amsterdam, The Netherlands, 1995.
44. Grinberg, M. *Flask Web Development: Developing Web Applications with Python*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2018.
45. CSI Piemonte. Last Updated: 2020-11-19. Available online: <https://www.csipiemonte.it/web/it> (accessed on 6 December 2020).
46. Gazzetta Ufficiale. DECRETO DEL PRESIDENTE DELLA REPUBBLICA 26 agosto 1993, n. 412, 1993. Available online: <https://www.gazzettaufficiale.it/eli/id/1993/10/14/093G0451/sg> (accessed on 6 December 2020).
47. Dati provenienti dal Geoportale della Città di Torino. Toponomastica-Dataset-CKAN. Last Updated: 2018-05-08. 2018. Available online: [https://sciamlab.com/opendatahub/dataset/c\\_l219\\_260](https://sciamlab.com/opendatahub/dataset/c_l219_260) (accessed on 6 December 2020).

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).