Exploring energy performance certificates through visualization

# Exploring energy performance certificates through visualization

Tania Cerquitelli*, Evelina Di Corso*, Stefano Proto*, Alfonso Capozzoli†, Fabio Bellotti*,

Maria G. Cassese*, Elena Baralis*, Marco Mellia‡, Silvia Casagrande§, Martina Tamburini§

* Department of Control and Computer engineering, Politecnico di Torino, Torino, Italy
† Department of Energy, Politecnico di Torino, Torino, Italy
‡ Department of Electronics and Telecommunications, Politecnico di Torino, Torino, Italy
§ Edison Spa, Torino, Italy

* † ‡ name.surname@polito.it
§ name.surname@edison.it

## ABSTRACT

Energy Performance Certificates (EPCs) provide interesting information on the standard-based calculation of energy performance, thermo-physical and geometrical related properties of a building. Because of the volume of available data (issued as open data) and the heterogeneity of the attributes, the exploration of these energy-related data collection is challenging. This paper presents INDICE (INformative DynamiC dashboard Engine), a new data visualization framework able to automatically explore large collections of EPCs. INDICE explores EPCs through both querying and analytics tasks, and intuitively presents the output through informative dashboards. The latter include dynamic and interactive maps along with different informative charts allowing different stakeholders (e.g., domain and non-domain expert users) to explore and interpret the extracted knowledge at different spatial granularity levels. The objective of INDICE is to create energy maps useful for the characterization of the energy performance of buildings located in different areas. The experimental evaluation, performed on a real set of EPCs related to a major Italian region in the North West of Italy, demonstrates the effectiveness of INDICE in exploring an EPC dataset through different data and knowledge visualization techniques.

## 1 INTRODUCTION

Nowadays large volumes of energy-related data are continuously collected in different domains. To reduce wasteful energy consumption, several orthogonal applications (e.g., buildings, IoT-based devices, wireless networks) increased their policy priority on energy efficiency. According to the U.S. Department of Energy, in industrialized countries more than 40% of total energy is consumed in buildings [14]. In the last few years many efforts have been devoted to improve building energy efficiency with different final goals: (i) facilitating proactive energy-saving services [32], (ii) characterizing data streams of energy consumption of individual residential consumers in buildings [5–7], (iii) characterizing heating energy demand through the analysis of energy performance certificates of buildings [4, 9, 11], and (iv) reducing emissions and energy consumption for buildings [20].



To enhance the effectiveness of data and knowledge exploration, a variety of data visualization techniques have been proposed. In [22, 23, 26] the authors exploited choropleth maps to analyze the energy consumption and the electricity consumption per unit area, respectively. Instead, in [21], the authors used dynamic simulations of building energy consumption and building information to develop urban energy maps with high spatial resolutions. However, all the above works proposed static maps to analyze the average values of some features of interest. The exploitation of dynamic and navigable maps tailored to the analysis of energy-related data has not been proposed so far. The authors in [24] propose an interactive 3D visualization to analyze the Linking Open Data (LOD) cloud adopting the metaphor of urban area. The visualization is interactive, meaning that the user can enlarge any part of the model, modify the perspective, change the shape of the buildings and their positioning, view all the connections or only those belonging to a specific data set. A parallel research effort has been devoted to explore and summarize geolocated time series data through maps [8]. Moreover, a great research effort has been done in [17], in which the authors propose a city energy model based on the requests and need for visualization from a group of energy consultants. Their proposed model offers stakeholders a powerful tool for evaluating both the current state and future scenarios.

This paper presents INDICE (INformative DynamiC dashboard Engine), a data visualization framework generating interactive and navigable dashboards through the analysis of a set of Energy Performance Certificates (EPCs). An EPC is a legal requirement when constructing, selling or renting a building, and it provides interesting information on the calculated standard energy performance, thermo-physical and geometrical properties of existing buildings. The multi-tiered framework INDICE has been proposed to effectively deal with large collection of EPCs. With respect to the other works, our framework brings together many different analysis techniques to help non-expert users make sense of Energy Performance Certificates. Indeed, after a pre-processing step, cluster analysis allows discovering groups of EPCs with similar features. To summarize the energy performance of buildings at different granularities, INDICE generates informative dashboards tailored to different energy stakeholders, combining both a rich set of interesting knowledge and ease of use.

The proposed informative dashboards exploit different kinds of energy maps to show data and knowledge at different spatial

granularity levels. The proposed visualization techniques allow different energy stakeholders to easily capture the high-level overview of heating energy demand at a city level, and drill-down the knowledge to the single apartment. Moreover, in order to analyze the energy efficiency of different buildings through the most interesting attributes under analysis, INDICE includes cluster-markers which dealing with the problem of representing multiple variables at the same time.

As a case study, a real collection of EPCs related to a major Italian region, in North West Italy, was analyzed. Preliminary experimental results show that the proposed approach is effective in visualizing a manageable set of human-readable knowledge for each end-user thought dynamic and interactive maps.

The next sections of the paper are organized as follows. Section 2 introduces an overview of the INDICE system with a thorough description of its main building blocks. Section 3 discusses the preliminary experimental results obtained on a real data collection and Section 4 draws conclusions and presents the future development of this work.

## 2 THE INDICE ANALYTICS SYSTEM

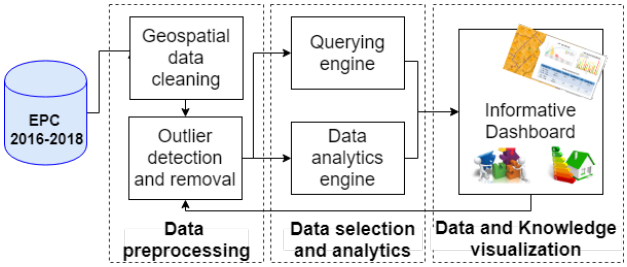


**Figure 1: The** INDICE **framework**

INDICE (INformative DynamiC dashboard Engine) has been tailored to analyze any collection of EPCs. The analysis of this kind of data is challenging, due to the large number of attributes characterizing each energy performance certificate. The exploitation of this high dimensional data is burdensome due to the high variability and dimensionality of data. INDICE combines different techniques to effectively visualize a rich set of knowledge items for a variety of energy stakeholders. The overall architecture is shown in Figure 1. INDICE includes three main building blocks, each one addressing one of the main steps of the knowledge-extraction process: (i) *Data pre-processing*, (ii) *Data selection and analytics*, and (iii) *Data and Knowledge visualization*. In the following, a detailed description of each building block is given.

### 2.1 Data Pre-processing

The INDICE pre-processing phase aims at smoothing the effect of possibly unreliable data. It performs two tasks which have been proved to be crucial in real-world geospatial data: (i) *geospatial coordinates cleaning* and (ii) *outlier detection and removal*.

#### 2.1.1 *Geospatial data cleaning.*

This pre-processing step is crucial when the final aim is to display data and knowledge through maps. INDICE includes an ad-hoc strategy to clean geospatial attributes, including address, house number, ZIP Code, latitude and longitude. Since the address attribute is usually collected as a free text field, it often contains numerous typos and input errors, which require careful analysis to be correctly fixed. To clean the above-mentioned attributes, INDICE includes a multi-step algorithm to correctly reconstruct and correct the wrong information. Specifically, it compares the available addresses with a referenced street map that is usually available for each city. The referenced street map should contain all the detailed information on streets, including street names, house numbers, ZIP Code and geolocation (i.e., latitude and longitude). Given a city under analysis, INDICE automatically downloads the referenced street map if it is available online.

The referenced street map is exploited by INDICE to verify the reliability of the addresses in the dataset under analysis to correct errors in the address field and at the same time reconstruct missing or incorrect information in the attributes ZIP Code, house address, latitude and longitude. Specifically, the developed algorithm compares each string in the dataset with the ones in the referenced street map. For each couple of addresses Levenshtein distance [19] is computed to evaluate the similarity between two character strings, in terms of the minimum number of modifications (insertions, deletions and substitutions) necessary to transform the first string into the second one. The similarity computed from Levenshtein distance takes values in the range [0-1], where 0 indicates total dissimilarity and 1 equality of the compared strings. Given a user-defined threshold $\phi$, the referenced address (the most similar to the address under analysis) replaces the original one if Levenshtein similarity between the two addresses is greater than or equal to $\phi$. When the association to a referenced address is not possible, i.e., Levenshtein similarities are below $\phi$, a geocoding request is sent via the Google Geocoding APIs[1]. The latter is a reliable service providing a textual address to reconstruct the whole address in a consistent way. However, INDICE exploits the Google Geocoding service only when the association cannot be resolved through the referenced street map due to a limit on the number of free requests.

#### 2.1.2 *Outlier detection and removal.*

An outlier is an extreme value that deviates from other observations on data. It may occur either when the collected value does not fit the model under study or when some error happens during the data collection phase. To address this issue, INDICE exploits three approaches: (i) univariate outlier detection, (ii) mixed univariate analysis, and (iii) multivariate outlier detection. Independently of the above adopted strategies, values labelled as outliers are not considered in the subsequent steps of analysis.

**Univariate outlier.** INDICE integrates three methodologies to automatically detect outliers and remove them for the subsequent analytics steps: (i) the graphic boxplot method, (ii) the parametric generalized Extreme Studentized Deviate (gESD) method [27] and (iii) the non-parametric Median Absolute Deviation (MAD) [15]. The **boxplot** [31] (aka whiskers plot) is a convenient way of visually displaying a data distribution through its quartiles. The frequency distribution of each variable is summed up through a few numbers (i.e. median, quartiles, min and max values). The median summarizes the central tendency of the distribution, while the quartiles give an indication of the variability through the interquartile difference. The minimum and maximum values provide not only information about extremes but also on the possible presence of data with abnormal characteristics w.r.t. the other points, plotting them individually. For each variable, the analyst can manually remove the outliers (i.e., the values smaller and greater than the minimum and the maximum) through value filters.

[1] https://developers.google.com/maps/documentation/geocoding/intro

The **gESD method** [27] is used to detect one or more outliers in a univariate data set. This test needs a parameter which is the upper bound on the number of potential outliers. INDICE tests the null hypothesis that the data has no outliers versus the alternative hypothesis that there are at most $k$ outliers (for some user specified value of $k$). Given the upper bound, $k$, the gESD test essentially performs $k$ separate tests: a test for one outlier, a test for two outliers, and so on up to $k$ outliers. In INDICE the number of outliers is determined by finding the largest value $r$ (with $r \leq k$), such that the corresponding test gives a value higher than the critical one.
Lastly, in statistics the **MAD method**[15] is a robust measure of the variability of a univariate sample of quantitative data. Calculating the MAD is straightforward, as it only involves finding the median of absolute deviations from the median. It is calculated by taking the absolute difference between each point and the corresponding median, and then calculating the median of those differences. As proposed in [16], INDICE uses the score of 3.5 as cut-off value. This means that every point with a score above 3.5 is considered an outlier.

The users can exploit all the different univariate methodologies and/or choose the most suitable one. If a non-expert user does not know how to deal with these outlier detection techniques, she can use default configurations, as described below.

**Expert-driven univariate analysis.** Because some non-expert users may be interested in analysing EPC collections, INDICE suggests the univariate outlier detection method mostly used by domain experts in the past interactions with INDICE. Specifically, by collecting and storing expert user (e.g., energy scientists) INDICE configurations, the non-expert users can receive interesting and effective suggestions to properly deal with noisy data. In the current version of INDICE, only relevant attributes describing the building thermo-physical characteristics (e.g., *Aspect Ratio*, *Average U-value of the vertical opaque envelope* and *Average U-value of the windows*) and the efficiency of the heating subsystems (e.g., *Distribution Subsystem Efficiency* and *Generation Subsystem Efficiency*) have been considered. In this way, if a non-expert user does not know which univariate analysis technique should be used, she can use a configuration adopted by previous INDICE expert users, since their choices are automatically stored as default configurations for non-expert users.

**Multivariate outlier detection.** For the multivariate outlier detection, INDICE integrates the DBSCAN algorithm (Density-Based Spatial Clustering of Application with Noise) [12] to automatically identify outliers. Specifically, DBSCAN detects clusters based on a density reachability concept, where clusters with higher-density regions are separated by lower-density regions. DBSCAN requires two user-defined parameters (i.e., *minPoints* and *Epsilon*). To properly specify these input parameters INDICE plots the k-distance graph and automatically estimates a good value for each parameter. As proposed in [10], INDICE runs several times the k-distance plot for different values of *minPoints*, and selects *minPoints* when the curve stabilises, and *Espilon* as the elbow point of the stable curve.

## 2.2 Data selection and analytics

The knowledge visualization step is preceded by a data selection and analytics phase. Since each energy performance certificate includes a large number of features characterized by a great variability, in order to extract accessible knowledge and implement data mining algorithms (e.g., cluster analysis, association rules) data have to be properly transformed and peculiar attributes have to be selected. Several techniques have been used to reduce the complexity of the datasets under analysis and discover effective and hidden knowledge, interesting and readable by all the different stakeholders involved in the analysis. This component includes two innovative engines: (i) the query engine and (ii) the data analytics engine.

### 2.2.1 *Querying engine.*

To select and explore the dataset under analysis, INDICE implements a query engine that lets the user focus on the single attributes of the energy performance certificates. Possible stakeholders may be citizens, public administration and energy scientists. Each of them could be interested in different characteristics of the dataset under analysis. For each stakeholder, INDICE produces the best possible representation to highlight the main interesting facets of the results. Citizens could be interested in the energy analysis of the buildings related to a specific area of the city, or in the geometric features that characterize the buildings belonging to the same intended use. The citizens may want to discover areas of the city with more performing buildings, to buy a flat that performs well in terms of energy efficiency. The public administration may be instead being interested in identifying areas where to promote and invest for energy renovations. Energy scientists could use INDICE to explore and characterize through supervised and unsupervised techniques groups of building with similar properties to perform benchmarking analysis. Based on the target of each stakeholder, the system is able to automatically propose to the specific end-user an optimal set of interesting reports and graphical representations, with the possibility to set manually the subset of features and parameters for the queries to which she is interested in.

### 2.2.2 *Data analytics engine.*

To extract meaningful and interesting knowledge items from data, INDICE includes different supervised and exploratory algorithms to automatically analyze feature subsets. INDICE integrates the K-means clustering algorithm [18] to create groups of buildings with similar thermo-physical and energy properties, and association rule mining [1] to extract interesting correlations among features.

**K-means algorithm.** The partitional K-means cluster algorithm [18] is exploited by INDICE to identify groups of EPCs characterized by similar properties. To measure the similarity between EPCs, the Euclidean distance is computed. The K-means algorithm, which is the most popular clustering algorithm, divides the input dataset into $K$ groups, where $K$ is defined a-priori. The average of all the energy certificates in each cluster represents the centroid (representative point) of each group of buildings. First, the algorithm chooses randomly K initial centroids. Then, each point is assigned to the closest centroid and the centroids are recalculated. The previous steps are repeated until the centroids no longer change. K-means is able to identify a good cluster set in a limited computational time. INDICE analyses the trend of the SSE (saReadum of squared error) quality index to evaluate the cluster cohesion [30] and automatically identify possible good $K$ values. The SSE is computed as the total sum of squared errors for all objects in the collection, where for each object the error is computed as the squared distance from the closest centroid. As done in [30], in INDICE the $K$ value is chosen as the point where the marginal decrease in the SSE curve is maximized (aka elbow approach).

**Association rules**. One of the most powerful exploratory techniques in data mining aiming at finding interesting correlations among data is represented by association rule discovery [1]. An association rule is expressed in the form $A \rightarrow B$, where $A$ and $B$ are disjoint and non-empty itemsets, (i.e., $A \cap B = \emptyset$). A is also called rule antecedent and B rule consequent. Since association rules extraction operates on a transactional dataset of categorical attributes, a discretization step is needed to convert the original continuously-valued measurements into categorical bins. The discretization adopted in INDICE are described in [11]. The used technique involves creating a decision CART (Classification And Regression Tree) [2] for each variable, using as response variable the annual primary energy demand normalized on the floor area. The tree splits are used as bins in the discretization process. To select only a subset of interesting rules, constraints on various goodness measures are used. INDICE includes four well-known quality indices: i) *support*, ii) *confidence*, iii) *lift*, and iv) *conviction*. The *rule support* is the percentage of transactions that contain both antecedent and consequent; *confidence* is the conditional probability that the consequent is true under the condition of the antecedent; *lift* [30] measures the correlation between the antecedent and the consequent; *conviction* [3] measures the degree of implication of a rule. Default thresholds are set by INDICE however the end-user could change the default values to analyze at different granularity level the extracted rules.

## 2.3 Informative dashboard

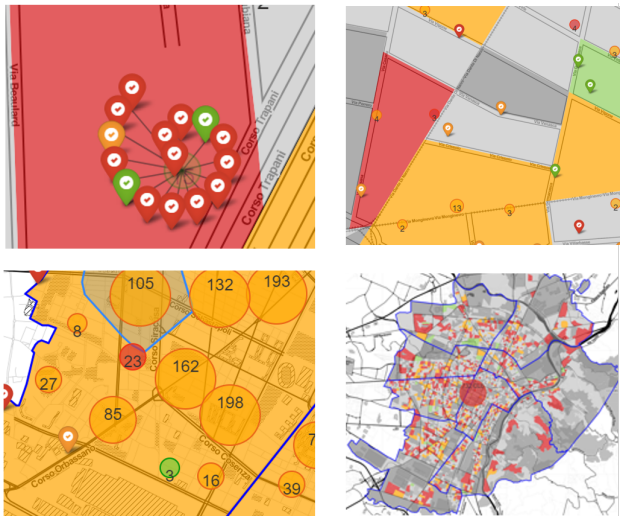


**Figure 2: Example of choropleth and scatter map at single certificate (Upper Left) and neighbourhood level (Upper Right), and Cluster-marker maps at district (Bottom Left) and city levels (Bottom Right).**

The aim of this component is to visualize and make the information and the extracted knowledge easy to be interpreted at different levels of detail. To this extent, INDICE includes interactive and navigable dashboards tailored to different use cases, providing both domain specific information and high-level energy demand overviews. Indeed, the dashboards can be customized for each end-user, providing deep targeted knowledge for domain experts and human-readable informative contents for non-expert users. Besides displaying charts and diagrams, which are typical of statistics and generally difficult to interpret, since the geolocalized EPC data lend themselves very well to be visualized on maps, INDICE proposes several techniques to explore and visualize the knowledge extracted from EPCs.
The dashboards include (i) geospatial maps, including traditional maps as choropleth and scatter maps and a new type of map named *cluster-marker map*, (ii) frequency distribution plots, (iii) association rules, and (iv) correlation matrices. These visualization techniques are jointly exploited by INDICE to graphically show the extracted knowledge at different spatial granularity levels such as city, district, neighbourhood, or housing unit (e.g., certificates belonging to the same building).

**Geospatial maps.** In INDICE, three geospatial maps have been integrated: (i) *choropleth maps*, (ii) *scatter maps*, and (iii) *cluster-marker maps*. These energy maps are related to each other, as each user can switch from one view to another, simply by changing the analysis zoom (i.e., drill down in the energy map) or introducing the knowledge of the cluster-markers. In **choropleth maps** each area (at different zoom levels) is colored according to the average value of the considered variable for the area under analysis. The **scatter maps** report a point and its corresponding value for each EPC (and so residential unit) contained in the selected area. **Cluster-marker maps**, similarly to the choropleth maps, aggregate multiple certificates coloring the dynamic markers according to the average of the values of the aggregated points. While the first two geospatial maps (i.e., choropleth and scatter maps) are useful for analyzing single variables, the cluster-marker visualization faces the problem of representing multiple variables at the same time. Specifically, exploring a single variable at coarse granularity levels could lead to flat and poor representative maps. To this extent, INDICE includes cluster-markers to introduce a new feature to the maps, in order to analyze the energy efficiency of several buildings through various attributes. The cardinality of the corresponding cluster affects the size of the marker and is reported inside the marker. These maps have been used together, ensuring in a single solution different levels of detail depending on the zoom degree selected by the user. Figure 2 shows examples of analysis results at different granularity levels, visualizing various information features on the maps. In the upper part of Figure 2, a set of attributes (i.e., the *Average U-value of the vertical opaque envelope* and the *Average U-value of the windows*, see Section 3 for further attribute details) extracted from the EPCs by means of the querying engine has been displayed. The choropleth map shows the average value of the attributes for the selected area together with the scatter marker of each single point, visualized at neighbourhood and housing unit zoom levels, respectively. The users can navigate the map and check the attribute values for each certificate by clicking on the markers. In the bottom part of Figure 2, the information obtained through the data analytics engine (e.g., the identification of the areas characterized by lower and medium energy performances) has been visualized at district (Left) and city (Right) levels. The cluster-markers show the cardinality of each cluster, together with the average value of an independent response variable chosen in the analytic process.

**Frequency distribution plots.** For a given area, the frequency distributions (e.g., quartiles or deciles) of the features selected for the visualization task are reported. A frequency distribution of data can be shown in a table or graph/diagrams. Some common methods include frequency tables, histograms or bar charts. These distributions can refer to single attributes or to aggregate information extracted from the analytic task, hence to groups of

similar certificates according to the subsets of attributes selected for the analysis. INDICE provides a setting panel to select one or more distribution visualizations, including the description of the main statistical indices. For numeric data, INDICE includes count, mean, standard deviation and the three quartiles (i.e., median, first and third quartiles), while for categorical attributes, the count, the most common value's frequency (i.e., mode) and the top-k frequent values are reported. The end-user can select a response variable against which to color the attribute distributions.

**Association rules.** INDICE discovers correlations in terms of association rules. However, to ease the manual inspection of the most interesting correlations, INDICE defines templates to characterize the attributes and represent the association rules using a tabular visualization. By sorting on quality indices, only the top-k rules that satisfy all constraints may be displayed. Rules can be extracted at different granularity levels, e.g., for each city, neighbourhood or downstream of the clustering algorithm.

**Correlation matrices.** To reduce the complexity of the analysis and remove correlated attributes from the analytic process, INDICE proposes correlation matrices to analyze the dependence between variables. For each pair of numerical attributes $X$ and $Y$, the framework computes the Pearson correlation coefficient [28], defined as $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ where $cov(X,Y)$ is the covariance between $X$ and $Y$, $\sigma_X$ is the standard deviation of $X$ and analogously $\sigma_Y$ for $Y$. Each coefficient value is translated into a gray level in the black-and-white scale to represent the correlation intensity in a plot matrix. When the selected set of attributes has no evident linear correlation, it is eligible for the analytic task.

## 3 PRELIMINARY EXPERIMENTAL RESULTS

INDICE has been experimentally evaluated on a real collection of building energy performance certificates. The EPCs are issued in the years between 2016 and 2018 for buildings and flats located in Piedmont, a major Italian region. This dataset has been collected and openly released by CSI Piemonte (the Information System Consortium)[2] and regulated by the Piedmont Region authority (Sustainable Energy Development Sector). The dataset includes approximately 25000 energy certificates, each one characterized by 132 features, including energy and thermo-physical attributes, divided into 89 categorical attributes and 43 quantitative attributes.

INDICE has been developed in Python [29], including the *scikit-learn* library [25] (for the analytic tasks) and *folium* library [13] (for visualization purposes).

### 3.1 Case study

To evaluate the effectiveness of INDICE, we focus on a case study having as stakeholder the public administration (PA). The results are obtained by tailoring the analysis to the city of Turin and selecting the EPCs related to the housing units of type E.1.1 (buildings used as permanent residence). To clean the geospatial coordinates, in the specific address, house number, ZIP Code, latitude and longitude for each EPC, INDICE applies the algorithm proposed in Section 2. This algorithm compares the addresses in the EPC dataset and the addresses in an open dataset[3] provided by the municipality of Turin, containing the city roads, with street names, house numbers, ZIP Code and geolocation (i.e., (latitude, longitude)). This database was used to verify the reliability of the addresses in our dataset. In our case study, if

[2]http://www.csipiemonte.it/web/it/

[3]https://www.sciamlab.com/opendatahub/dataset/c_l219_260

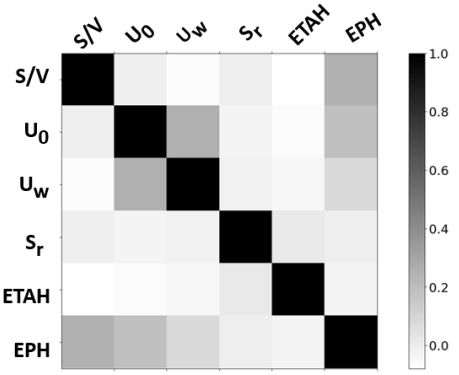


**Figure 3: Correlation matrix between pairs of numerical attributes**

the PA user is interested in discovering which areas of a city are more energy consuming and which are more efficient, she could select the following subset of attributes, which characterize the thermo-physical properties of each building: *Aspect Ratio* (S/V), *Average U-value of the vertical opaque envelope* ($U_o$), *Average U-value of the windows* ($U_w$), *Heat surface* ($S_r$) and *Average global efficiency for space heating* (ETAH). The Aspect Ratio represents the geometric shape of a building. $U_o$ and $U_w$ measure the heat loss through the opaque and the transparent elements of the building, respectively. The lower the thermal transmittance of the building envelope, the lower the heat flow that is transmitted through the elements themselves. The Heat surface corresponds to the heated floor area. Lastly, the ETAH index takes into account all the thermal losses of each subsystem, including the generation, distribution, emission and control subsystems. The PA user may be interested in discovering groups of buildings with homogeneous thermo-physical properties. To address this task the K-means clustering algorithm can be applied.

Before clustering, the correlation between the considered numerical attributes is checked. In Figure 3, the correlation plot matrix between the considered attribute pairs is reported. Dark squares represent high linear correlation between the two variables, while light squares represent low correlation. All the variables considered in the analysis are weakly correlated (i.e., there is no evident linear association between variable pairs). Hence, the results obtained from the five attributes selected for the clustering phase (i.e., S/V, $U_o$, $U_w$, $S_r$ and ETAH) and the response variable *Normalized primary heating energy consumption* (EPH), allow the extraction of non-trivial knowledge from data. Figure 4 shows the results obtained by the data analytics engine for the features described above. From the charts reported in the dashboard, the analyst can explore the frequency distribution of a specific attribute, as the response variable EPH, or its distribution in the cluster set detected by INDICE. Moreover, interesting correlation rules[4] can be extracted and visualized using a tabular representation. In this way, every end user, independently of her expertise degree, can detect the attributes which influence most the energy performance of buildings and find out the geographical areas for which a certain set of rules apply. Driven

[4]The discretization used for the dynamic dashboard is as follows. 4 classes for the Average U-value of the windows (i.e., Low = [1.1, 2.05], medium = (2.05, 2.45], High = (2.45, 3.35] and Very high = (3.35, 5.5]); 3 classes for the Average U-value of vertical opaque envelope (i.e., Low = [0.15, 0.45], medium = (0.45, 0.65], High = (0.65, 1.1]; 3 classes for the Average global efficiency for space heating (i.e., Low = [0.20, 0.60], medium = (0.60, 0.80], High = (0.80, 1.1].

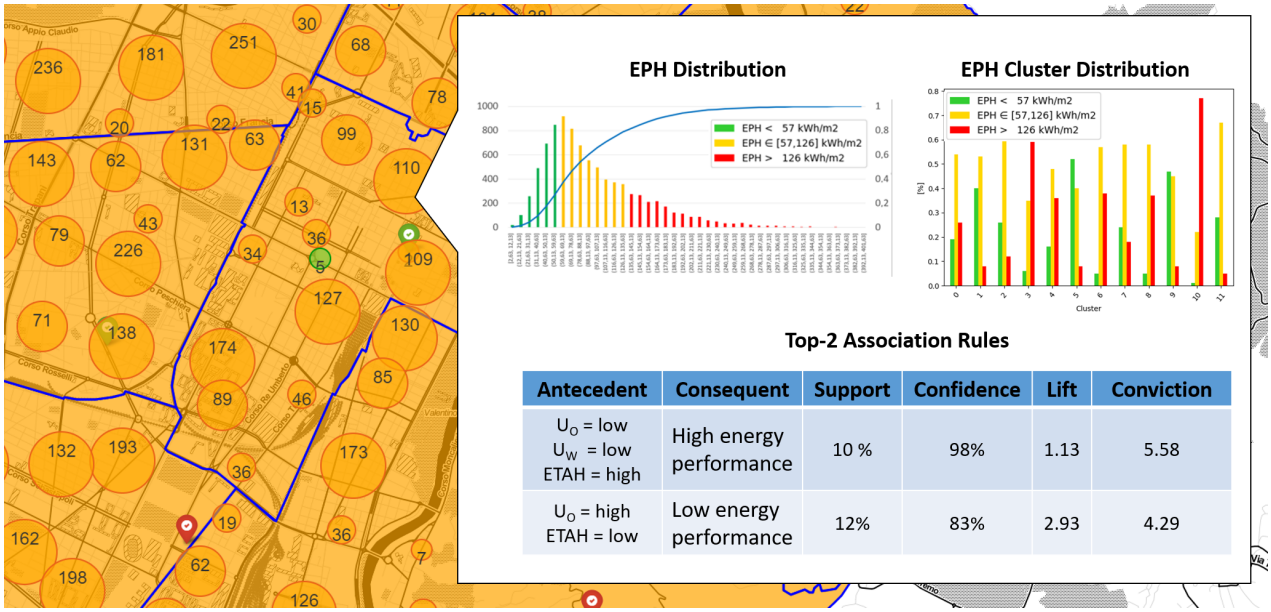


**Figure 4: Interactive dashboard visualizing, at district level, the result of the data analytics engine.**

by the extracted knowledge, the PA user may support and incentive renovation policies targeting specific low performance neighborhoods, or identifying groups of similar EPCs.

## 4 CONCLUSIONS AND FUTURE WORKS

This paper presents INDICE, a new data visualization framework that analyzes EPC collections at different granularity levels. After a preprocessing step, INDICE extracts interesting and hidden knowledge for different end-users. Informative dynamic dashboards have been presented to show useful information, at different geospatial levels and with enriched map representations (e.g., the cluster-marker map).

As future work we plan to integrate in INDICE other analytics techniques (both supervised and unsupervised) to provide a more flexible and enhanced analysis. Furthermore, the analysis process should be empowered by an automatic tool suggesting appropriate analysis configurations for the considered datasets. To this aim, we are currently planning to release our framework INDICE in order to have real feed-backs from end-users (e.g., citizens, energy experts, public administration). In this way, we could improve the choices of the default configurations, but also include and integrate further representations to improve the visualization of the extracted knowledge.

## Acknowledgments

The research leading to these results has been supported by the SmartData@PoliTO center for Big Data and Machine Learning technologies.

The authors express their gratitude to Giovanni Nuvoli (Settore Sviluppo Energetico Sostenibile - Regione Piemonte) and to CSI Piemonte.

## REFERENCES

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. In *Acm sigmod record*. ACM, 207–216.

[2] Leo Breiman. 2017. *Classification and regression trees*. Routledge.

[3] Sergey Brin, Rajeev Motwani, Jeffrey D Ullman, and Shalom Tsur. 1997. Dynamic itemset counting and implication rules for market basket data. *Acm Sigmod Record* 26, 2 (1997), 255–264.

[4] Alfonso Capozzoli, Daniele Grassi, Marco Savino Piscitelli, and Gianluca Serale. 2015. Discovering Knowledge from a Residential Building Stock through Data Mining Analysis for Engineering Sustainability. *Energy Procedia* 83 (2015), 370 – 379. https://doi.org/10.1016/j.egypro.2015.12.212

[5] Tania Cerquitelli, Gianfranco Chicco, Evelina Di Corso, Francesco Ventura, Giuseppe Montesano, Mirko Armiento, Alicia Mateo González, and Andrea Veiga Santiago. 2018. Clustering-Based Assessment of Residential Consumers from Hourly-Metered Data. In *2018 International Conference on Smart Energy Systems and Technologies (SEST)*. IEEE, 1–6.

[6] Tania Cerquitelli, Gianfranco Chicco, Evelina Di Corso, Francesco Ventura, Giuseppe Montesano, Anita Del Pizzo, Alicia Mateo González, and Eduardo Martin Sobrino. 2018. Discovering electricity consumption over time for residential consumers through cluster analysis. In *2018 International Conference on Development and Application Systems (DAS)*. IEEE, 164–169.

[7] Tania Cerquitelli and Evelina Di Corso. 2016. Characterizing Thermal Energy Consumption through Exploratory Data Mining Algorithms.. In *EDBT/ICDT Workshops*.

[8] Georgios Chatzigeorgakidis, Dimitrios Skoutas, Kostas Patroumpas, Spiros Athanasiou, and Spiros Skiadopoulos. 2018. Map-Based Visual Exploration of Geolocated Time Series. In *Proceedings of the Workshops of the EDBT/ICDT 2018 Joint Conference (EDBT/ICDT 2018), Vienna, Austria, March 26, 2018*. 92–99. http://ceur-ws.org/Vol-2083/paper-14.pdf

[9] Giuliano Dall'O, Luca Sarto, Nicola Sanna, Valeria Tonetti, and Martina Ventura. 2015. On the use of an energy certification database to create indicators for energy planning purposes: Application in northern Italy. *Energy Policy* 85, C (2015), 207–217.

[10] Evelina Di Corso, Tania Cerquitelli, and Daniele Apiletti. 2018. METATECH: METeorological Data Analysis for Thermal Energy CHaracterization by Means of Self-Learning Transparent Models. *Energies* 11, 6 (2018), 1336.

[11] Evelina Di Corso, Tania Cerquitelli, Marco Savino Piscitelli, and Alfonso Capozzoli. 2017. Exploring Energy Certificates of Buildings through Unsupervised Data Mining Techniques. In *Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2017 IEEE International Conference on*. IEEE, 991–998.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*. 226–231.

[13] Filipe and Martin Journois at all. 2018. python-visualization/folium: v0.6.0. (Aug. 2018). https://doi.org/10.5281/zenodo.1344457

[14] Xiaohong Guan, Zhanbo Xu, and Qing-Shan Jia. 2010. Energy-efficient buildings facilitated by microgrid. *IEEE Transactions on smart grid* 1, 3 (2010), 243–252.

[15] Frank R Hampel. 1974. The influence curve and its role in robust estimation. *Journal of the american statistical association* 69, 346 (1974), 383–393.

[16] Boris Iglewicz and David Caster Hoaglin. 1993. *How to detect and handle outliers*. Vol. 16. Asq Press.

[17] Tim Johansson, Mattias Vesterlund, Thomas Olofsson, and Jan Dahl. 2016. Energy performance certificates and 3-dimensional city models as a means to reach national targets–A case study of the city of Kiruna. *Energy Conversion and Management* 116 (2016), 42–57.

[18] B.-H. Juang and L.R. Rabiner. 1990. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing* 38, 9 (Sep 1990), 1639–1641.

[19] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*. 707–710.

[20] Xue Li, Lanshun Nie, and Shuo Chen. 2014. Approximate Dynamic Programming Based Data Center Resource Dynamic Scheduling for Energy Optimization. In *IEEE iThings/GreenCom/CPSCom 2014, Taipei, Taiwan, September 1-3, 2014*. 494–501.

[21] Feng-Yi Lin, Tzu-Ping Lin, and Ruey-Lung Hwang. 2017. Using geospatial information and building energy simulation to construct urban residential energy use map with high resolution for Taiwan cities. *Energy and Buildings* 157 (2017), 166–175.

[22] Sara Torabi Moghadam, Patrizia Lombardi, and Guglielmina Mutani. 2017. A mixed methodology for defining a new spatial decision analysis towards low carbon cities. *Procedia Engineering* 198 (2017), 375–385.

[23] Y Olivo, A Hamidi, and P Ramamurthy. 2017. Spatiotemporal variability in building energy use in New York City. *Energy* 141 (2017), 1393–1401.

[24] Maria-Evangelia Papadaki, Panagiotis Papadakos, Michalis Mountantonakis, and Yannis Tzitzikas. 2018. An Interactive 3D Visualization for the LOD Cloud. In *Proceedings of the Workshops of the EDBT/ICDT 2018 Joint Conference (EDBT/ICDT 2018), Vienna, Austria, March 26, 2018*. 100–103. http://ceur-ws.org/Vol-2083/paper-15.pdf

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[26] Iraci Miranda Pereira and Eleonora Sad de Assis. 2013. Urban energy consumption mapping for energy management. *Energy Policy* 59 (2013), 257–269.

[27] Bernard Rosner. 1983. Percentage points for a generalized ESD many-outlier procedure. *Technometrics* 25, 2 (1983), 165–172.

[28] Sheldon M Ross. 2014. *Introduction to probability models*. Academic press.

[29] Guido Rossum. 1995. *Python Reference Manual*. Technical Report. Amsterdam, The Netherlands, The Netherlands.

[30] Pang-Ning Tan et al. 2007. *Introduction to data mining*. Pearson Education India.

[31] John W Tukey. 1977. Box-and-whisker plots. *Exploratory data analysis* (1977), 39–43.

[32] Chao-Lin Wu, Wei-Chen Chen, Yi-Show Tseng, Li-Chen Fu, and Ching-Hu Lu. 2014. Anticipatory Reasoning for a Proactive Context-Aware Energy Saving System. In *IEEE iThings/GreenCom/CPSCom 2014, Taipei, Taiwan, September 1-3, 2014*. 228–234.