Multidimensional analyses tools for energy efficiency in large building stocks

Dario Cottafava
Physics Department
University of Turin, Italy
e–mail: dario.cottafava@unito.it

P. Gambino[*], A. Tartaglino[**], M. Baricco[***]
Physics Department[*], Energy Manager[**], Chemistry Department[***]
University of Turin, Turin, Italy
e–mail: paolo.gambino@unito.it[*], andrea.tartaglino@unito.it[**], marcello.baricco@unito.it[***]

ABSTRACT

Energy Indices and Targets for large building stocks are a still complex field of debate. While various indices have been suggested, for instance based on Heating Degree Days or energy consumption per square meter / user, there is no generally recognized index for large building stocks, because of several factors which complicate their analysis. The heterogeneity in building functions and structures, number of users and opening hours, for instance, directly affects the energy consumption.

Our approach is based on multidimensional analysis of various Energy Efficiency Indices (EEI), as energy consumption per square meter or the day / night energy efficiency index, computed from simple data available for any building: monthly energy consumptions, total area, and building functions. Starting from available data for the University of Turin, composed by over than 46 buildings, and thanks to Data Visualization Javascript libraries, an interactive web application has been developed. In this paper, four different types of multidimensional plots, based on the scatter method and the parallel coordinates method, will be discussed. These approaches can be employed to form clusters of buildings and to identify specific thresholds for indices with respect to the functions of the buildings or other factors.

Where the most common and simplest indices actually fail, the setting of alert thresholds for the energy consumption of buildings with different functions can support the Energy Management in meeting efficiency targets. Results obtained for specific building types have been reported.

KEYWORDS

Energy Management, Energy Efficiency, Sustainability, Data Visualization, Clustering technique, Data Mining.

[*] Physics Department, Energy Working Group of UniTo Green Office
[**] Energy Manager, Energy Working Group of UniTo Green Office
[***] NIS Centre of Excellence and Chemistry Department, Energy Working Group of UniTo Green Office

INTRODUCTION

All over the world many programs have been promoted, adopted and enacted during the last decade to enhance building sustainability. These include, for instance, the 20-20-20 policy for the reduction of greenhouse gas (GHG) emissions by the European Union (EU) and the LEED program in the U.S. [1]. In fact, the primary energy consumption of residential and service buildings is one of the main sources of GHG emissions. Within the EU, the buildings energy consumption reaches about 40% of the total energy consumption [2,3], while in the U.S. residential and commercial buildings account for over 70% of the total electric energy consumption [4]. These values imply that buildings are a primary target for energy efficiency research. While most of the cited energy efficiency programs focus on the renovation of buildings, other factors strongly affect the energy consumption and correlated GHG emissions, such as occupants habits and the management process [5].

In this paper we will discuss a general approach, focused on the management process, the identification of anomalies and the definition of energy consumption thresholds, depending on the functions of the buildings. The approach is mainly based on data visualization techniques and focused on the energy consumption dataset of a large building stock.

Necessity of energy targets

The need for precise targets and indices to measure energy consumption in Universities, as well as their environmental impact, is widely recognized. It has also been recognized by the International Sustainable Campus Network (ISCN), founded in 2007. The ISCN is a no-profit, global forum with the aim to lead "*colleges, universities, and corporate campuses in exchanging information, ideas and best practices for achieving sustainable campus*". The ISCN is organized into four working groups in order to help colleges and universities fulfil their leadership role in creating a sustainable future. In particular, one of these groups, the "*Campus-Wide Planning and Target Setting*" group, is focused on the strategic planning of sustainability goals and targets [6].

As recognized by the ISCN, also in literature the necessity of energy targets and indices is broadly studied. Many papers focus on the energy efficiency indices (EEI) and dozens of indices have been proposed worldwide. Abu Bakar et al. [7] studied the EEI to compare buildings performance based on heating, ventilating and conditioning systems. Moghimi et al. [8] analysed EEI on commercial buildings based on the occupied air conditioning area. González et al. [9] proposed to compare the energy consumption with a reference building. The actual challenges focus on estimating, quantifying accurately and developing a relationship for EEI. To simplify, in fact, several researchers consider separately the different types of buildings to avoid unfair comparison because each building does not have the same function. [10].

The University of Turin's building stock. In this very general global framework, the University of Turin may be a particularly interesting case study in order to test indices and innovative approaches. In fact, the University of Turin (Unito) manages a large stock of about 120 buildings in the City of Turin and in the surrounding, for a total of over 800000 m^2. Unito has almost 70000 students, 2000 professors and 1800 administrative and technical staff. The annual energy cost is more than 10 million €, relatively to 23.5 GWh of electrical energy and 2.08 TOE of methane gas. Unito's building stock is quite heterogeneous: it includes libraries, research centers, administrative offices, museums, botanical gardens, hospitals and so on, with several distinctions in management. For instance, libraries and museums have a standard opening schedules, hospitals and research centers work 24/24h. In fact, generally, the opening schedule depends on the field of the department: the departments of humanities have typically a standard schedule, while departments of science have sometimes restricted access also during the weekend. Moreover, buildings in Unito's stock belong to different historical period with completely distinct architectural features: for example, some buildings within the city centre were built in the XVIth and the XVIIIth century (Rectorate - 1713, the Maths Department - 1675, Stemmi Palace – 1683) while newer buildings are from XX and XXI centuries (Palazzo Nuovo – 1966, Agrarian Campus – 1999, Campus Luigi Einaudi – 2012). This heterogeneity obviously affects any comparison of energy consumption for different buildings. A first achievement has been obtained making data available and developing online web tools [11].

A clusters approach. To improve energy management, as described in the following, this work aims to identify proper targets and to define thresholds in energy consumption for different classes of buildings, mostly depending on their function. Various research works underline the necessity of a clusters approach to analyse large bulding stock. For instance, Sonetti et al. [12] argued as some of the most widely recognized green rankings for universities lack of a precise analysis based on the typology of the universities or based on the building types which they are composed by. Moreover, a study of the National Bureau of Stastistics of China [13] highlighted as universities with medical college, polytechnic universities or megaversities, for instance, have energy consumption per floor area (per square meter) which cannot be ranked with the same criteria. In their work, Sonetti et al., proposed three clusterizations based on urban morphology, climate zones or university functions demonstrating how consumption strongly differs from buildings with different functions. On the contrary, several studies, as Mutani et al. [14], proposed very precise GIS analyses in order to characterize building thermal energy consumption at the urban scale and to visualize geolocalized clusters on maps depending on several factors, such as the surface to volume ratio. In fact, energy efficiency analyses are already widely known by engineers and architects. The major problem for energy efficiency studies on a large scale is that, generally, architectural analyses need many data or information (usually not available) and typically are very time-consuming for energy management.

The proposed process and approach may permit to explore the energy consumption dataset of a large building stock and to define clusters depending on the functions of the buildings in few steps, simply starting from energy bills and exploiting data visualization techniques. More precisely, in our work, we apply the very general approach described by Inselberg [15], the multidimensional detective, to energy efficiency analyses, in order to define clusters thresholds. Starting from the suggestions of Sonetti et al., we determine precise alert thresholds on energy consumption per square meter and, in order to improve the method, we also classify buildings based on the day/night energy efficiency index. The proposed approach aims to simplify the clusterization process with respect to GIS analyses and, meanwhile, to define precise rules to clusterize buildings depending on some indices as suggested by Sonetti.

The rest of the paper is structured as follows. In "Data visualization" section, some well-known data visualization tools will be described and a new application for energy efficiency indices will be suggested. In "Multidimensional detective approach" section, an intuitive approach based on visual cues will be presented. Finally, in "Results" and "Discussion" sections, the obtained results, based on an explorative multidimensional-detective approach, will be discussed. In the whole paper, we employ the dataset of the energy consumption of the Unito's building stock.

DATA VISUALIZATION

With the evolution of technology, during last decades a huge amount of data has become ubiquitous. Each specific dataset available today contains an enormous potential and possible information. In this framework, data visualization has a very important role to help researchers and data scientists to understand hidden and not straightforward information lying in any possible dataset. As stated by Card et al. [16] the most general definition of Information Visualization is: *"Information visualization is the use of computer-supported, interactive, visual representations of abstract data to amplify cognition".* In particular, data visualization has been defined as the process of using graphical presentation to transform complex data into visual insights, in order to provide to the user a qualitative information. For this purpose, the major aims of data visualization are, in fact, *presentation, confirmative and explorative analyses* [17].

The first cited aim (i.e. presentation) is obviously to present data and to help users understand hidden information. For this purpose, an interactive process, supported by data visualization techniques, is fundamental to allow users to redefine new insight, simply manipulating data in real time. Even if presentation is one of the most important goal for data visualization, the explorative and confirmative aspects can be the most relevant for sustainability studies. In data-driven studies, when faced with an unknown dataset, the first observation of an explorative analysis aims at possible structures and models, to identify trends and meaningful variables, to test preliminary hypotheses and to detect anomalies and outliers. Finally, data visualization, linked to clustering techniques, can be exploited for confirmative analyses in order to test hypotheses and theses [18]. In particular, this paper will focus on an explorative analysis of energy data related to the buildings of the University of Turin.

Thus, in order to extract the valuable information hidden in the data, each multidimensional multivariate dataset can be defined as a set of "*observations*" X where the ith element x_i is an array with m variables, $x_i = (x_{i1}, x_{i2}, \dots, x_{im})$. Thus, generally speaking, a multidimensional multivariate dataset X can be defined as a matrix:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

where m is the number of dimensions (attributes) and n is the number of "observations" (the size of the dataset) and x_{ij} is the value of the observation i with attribute j. For any multidimensional multivariate dataset, broadly speaking, data visualization techniques may be subdivided into:

- *Axis reconfiguration* maps each dimension to a different axis [19];
- *Dimensional embedding* exhibits a hierarchical visualization of the dataset [20];
- *Dimensional sub-setting* groups dimensions into subsets of 2-3 dimensions [21];
- *Dimensional reduction* maps large dimensional spaces into lower dimensional spaces [22].

Two of the most popular techniques for data visualization, i.e. the *"Parallel Coordinates*", an axis reconfiguration technique, and the *"Scatter Plot Matrix",* a dimensional sub-setting technique, will be discussed in this paper.

Tested variables and indices

Starting from the raw energy data, several energy efficiency indices (EEI) have been computed. The most general definition of an Energy Efficiency Index is [23]:

$$EEI = \frac{E}{\Lambda} \tag{1}$$

where $E =$ «Energy Consumption» and $\Lambda =$ «Factor related to the Energy Consumption». The factor E can be related to the Electrical Consumption E_{KWh} (KWh) or to the gas consumption E_{gas} (m^3 of gas), while the factor Λ can represent the surface of the building Λ_{m^2} (m^2), the number of users Λ_{users} (n° of users), the degree days Λ_{DDx} (degree days), where DD_x refers to several approaches for degree days (e.g. winter or summer), or to any other factors. Another useful energy efficiency index is based on monitoring the consumption during the week working hours and during the night / weekend / holiday. Computing the ratio of the energy consumption at night / weekend / holiday over the energy consumption during the working hours it is possible to obtain a preliminary index about the efficiency of the Heating, Ventilating and Conditioning (HVAC) systems and the lighting management as described in [11]. The "*night/day energy efficiency index*" for the electrical energy consumption for a whole year can be defined as:

$$EEI_{year,KWh,night/day} = \frac{1}{N}\sum_{i=1}^{N} EEI_{i,KWh,night/day} = \frac{1}{N}\sum_{i=1}^{N} \frac{E_{i,KWh,night}}{E_{i,KWh,day}} \tag{2}$$

where $E_{i,KWh,night}$ = «KWh during night / weekend / holiday», $E_{i,KWh,day}$ = «KWh during working hours within weeks» for month i and $N = 12$ is the number of months. This ratio simply reveals if the heating plants, the lightings and all other electrical loads has been properly set. In fact:

- if $EEI_{i,KWh,night/day} > 1 \Rightarrow E_{i,KWh,night} > E_{i,KWh,day}$
- if $EEI_{i,KWh,night/day} < 1 \Rightarrow E_{i,KWh,night} < E_{i,KWh,day}$

In conclusion, this ratio can quickly reveal inefficiencies in the lighting and heating schedules.

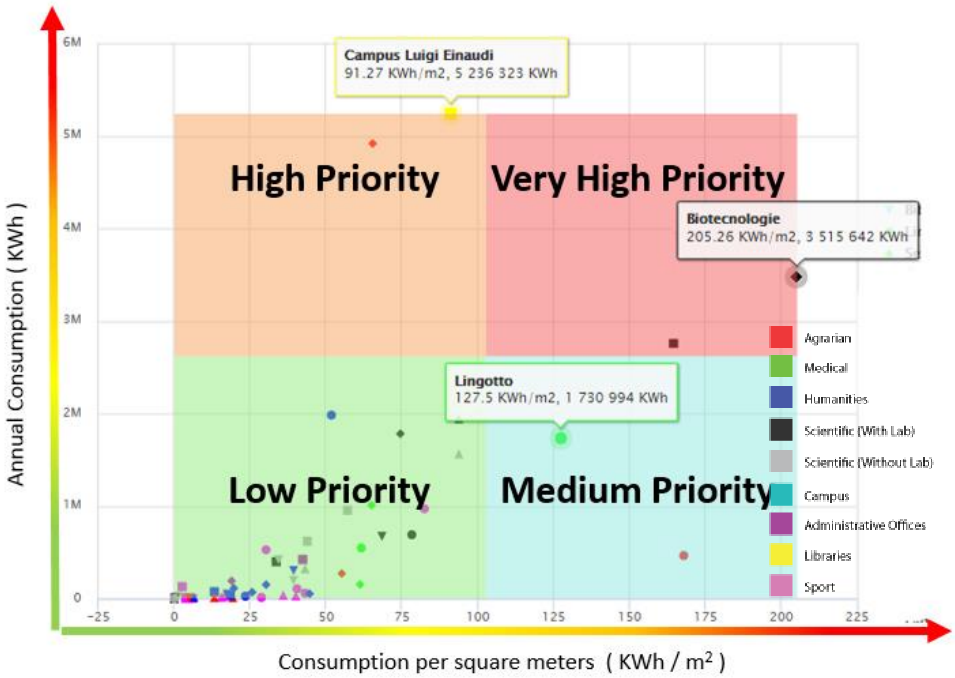

Figure 1. Graph represents the scatter method and the corresponding four priority quadrants. The x-axis corresponds to the electrical consumption per square meter (KWh / m^2) while the y-axis the annual electrical consumption (KWh).

Scatter method

The scatter method is a preliminary analysis method useful to identify the top priorities for energy auditing in a large building stock [24]. The scatter method is based on a very intuitive analysis. First, two dependent, or independent, energy related variables have to be chosen for the x-axis and the y-axis. Second, the graph has to be split into four different quadrants; in this way four different "*priority*" zones can be defined. Finally, all data points, one for each building, have to be plotted. As described in [11], one possible choice for the variables can be to plot on the x-axis the *normalized electrical annual consumption* per square meter (KWh / m^2) and on the y-axis the *total consumption during a year* (KWh). With this choice the x-axis represents a rough indicator of the efficiency of a building, while the y-axis represents its impact within the building stock. Consequently, the four quadrants represent different priorities: the top-right and the top-left quadrants represent the "very high priority" and "high priority", because of high total consumption, while the bottom-left and the bottom-right quadrants correspond to "low priority" and "medium priority", because of the low total consumption. Fig. 1. shows the application of this method to the selected set of buildings. For the sake of completeness, it must be stressed that the choice described above is only one over several possibilities. In fact, different variables can be analysed with the same method exploring possible correlations and identifying various ranking of priority for intervention.

Figure 2. Graph exhibits the scatter plot matrix, a collection of single scatter graphs, each one with a different couple of variables. The boxes in the diagonal have the same variable for the x and the y-axis.

Scatter Plot Matrix. The scatter plot matrix is one of the most popular data visualization tools. As described by Keller [25], the scatter plot matrix exhibits pairwise relationship among couples of variables. A matrix of single scatter plot reveals the relationships, as in a correlation matrix. Thus, the scatter plot matrix is simply a generalization of the scatter method where, within the same visualization, many couples of variables may be visualized all together. The method is shown in Fig. 2

Parallel coordinates method

Inselberg [26] developed the *Parallel coordinates* method and introduced it in order to visualize multidimensional datasets. Perpendicular to the x-axis, m copies of y-axis, $X_1, X_2, \ldots, X_m$, have to be set equidistant. Generally, the X_j axis lies at position $j-1$ for $j = 1, 2, \ldots, m-1$ perpendicular to the x-axis. Each axis represents one dimension of the m-dimensional space R^m. With this representation, for instance, the observation $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ is a polygonal line with vertices at $(j-1, x_{ij})$ on the X_j axis for $j = 1, 2, \ldots, m-1$. In general, a one-to-one correspondence is guaranteed between an observation x_i and a planar polygonal line. An example based on Unito's building stock is provided in Fig. 3. In this case, buiding type, electrical consumption per m^2, the ratio of electrical consumptions on night over day and the total annual consumption have been selected as dimensions.

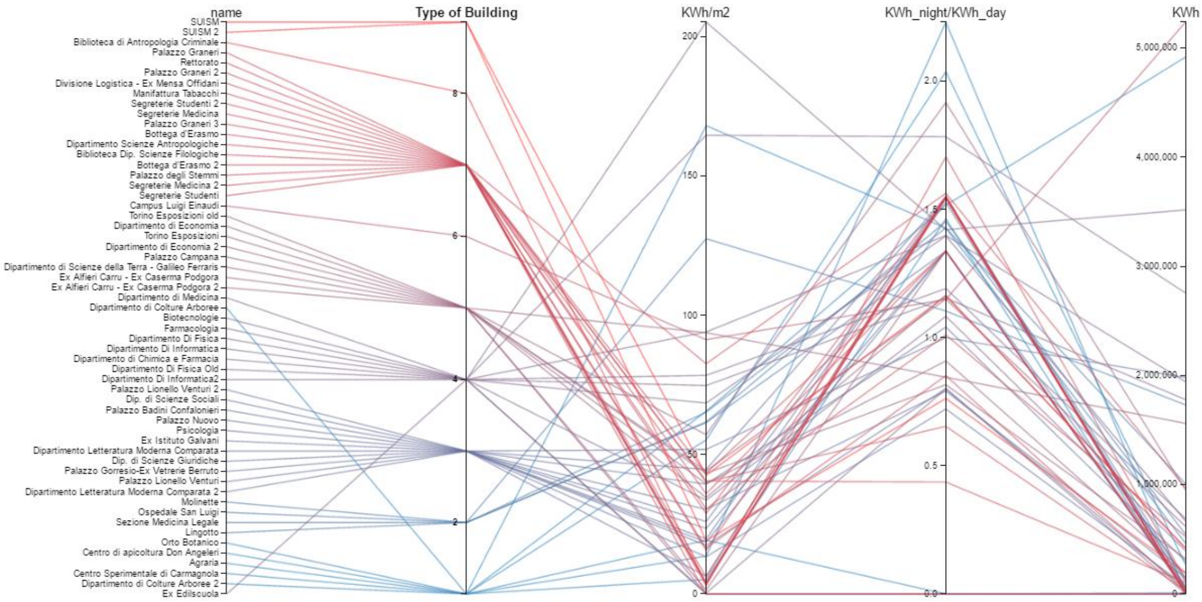

Figure 3. An example of parallel coordinates based on the Unito's building stock.. Each vertical axis represents a particular category / dimension of the multidimensional multivariate dataset (type of building, consumption per square meter, Day/night energy efficiency index, annual total consumption).

There exists one fundamental property called "Bumping the Boundaries" [23], exploited to clusterized data within a dataset, which states: "*A polygonal line which is in-between all the intermediate curves/envelopes represents an interior point of the hypersurface and all interior points can be found in this way. If the polygonal line crosses anyone of the intermediate curves it represents an exterior point"*. In particular, this property has been exploited to create the clusters in our approach. In fact, the data miner can select a single range within an axis (dimension), or multiple range simultaneously from different axes. In this way, all observations (buildings) lying in those particular intervals, might belong to a given subset (cluster) of *X*.

Historical trends. Parallel coordinates can be exploited in order to show the historical trend of a given variable (for instance the annual power consumption). In this specific case, the different dimensions for each observation represent the historical annual power consumption. Thus the real value x_{ij} represents the power consumption of the ith observation" (the ith building) of the jth dimension (the jth year).

MULTIDIMENSIONAL DETECTIVE APPROACH

As described by Inselberg [19] the explorative process consists in few general steps based on four tips and suggestions: "*Do not let the picture intimidate you*", "*Understand the objectives and use them to obtain visual cues*", "*Carefully scrutinize the picture*" and "*Test the assumptions*". The main aim of the visual cues is to help the researcher to focus on the interesting parts of the dataset. In fact, the visual cues approach can be exploited as a *pre-processor*, providing unique insights in order to define hypotheses and to reduce the computational time of clustering algorithms. Indeed, a n-points dataset has 2^n possible subsets (clusters). Each of these clusters can be the interesting one depending on the investigation field and the insight of the data scientist. Thus, the multidimensional detective approach can provide a tool to reduce this combinatorial explosion when no cues can be directly inferred from the dataset but only from additional knowledge (for instance from the energy management). In conclusion, the multidimensional detective approach can be exploited, for relatively small datasets. This analysis allows to identify outliers starting from a first raw clusterization, to recognize "junk variables" (dimensions) which have no meaning for the considered dataset and to define "raw boundaries" for each dimension, depending on the particular aspect that the data miner wants to analyse. Setting raw boundaries means to set a range, and alert thresholds, for any different cluster (e.g. building type) related to a valuable dimension j, (e.g. energy consumption per square meter) where the real value x_{ij} of an observation must lie. In other words, a minimum, and a maximum, value has to be set, such as $x_{min,j} \leq x_{ij} \leq x_{max,j}, \forall x_i \in O_k$ where O_k is the k-th subset of *X*.

In this work, a step-by-step process has been followed in order to apply multidimensional detective approach to the energy efficiency issues for a large building stock. Our approach is based on three main steps, enhancing the four tips given by Inselberg: define building types, test the assumptions and identify thresholds and outliers. As a first step, the data miner has to choose and to define building types, such as administrative offices, sport infrastructures, research centers and so on. In this first phase, the data miner can be helped from background knowledge of the energy management, can take advantage of some clustering algorithm techniques or try to identify building types from various visualizations. Once defined the functions of the buildings, a single building type has to be assigned to each building. Secondly, the data miner has to test the assumptions about alert thresholds for any given cluster. If there are few outliers and clear boundaries among the chosen clusters, the assumptions may be considered correct. Otherwise, the data scientist has to refine the assumptions. Finally, the data miner can define the boundaries and, then, identify outliers, which probably represent most inefficient buildings.

RESULTS

In this section, four approaches for various indices will be discussed. The exploited methods belong to the two families of data visualization techniques: the axis reconfiguration technique and the dimensional sub-setting technique. More in detail, the scatter plot method, within the dimensional sub-setting technique, has been employed for two main purposes: first, to identify the most inefficient buildings and to create macro-clusters of buildings depending on priorities, as described in section *Data visualization*, and second, to test the clusters described in the following.

Dataset description

All energy-related data, on a monthly basis, have been obtained directly from the energy bills of the University of Turin. In particular, the analysed dataset counts 46 buildings, 59 electricity meters and 77 methane gas meters. Each energy consumption can be split into three subsets: in fact, the Italian electricity market separates the consumption into three time slots, called F1, F2 and F3. More precisely, F1 corresponds to 8:00-19:00, Monday to Friday, F2 corresponds to 7:00-8:00 and to 19:00-23:00, from Monday to Friday, and 7:00-23:00 on Saturday, while F3 corresponds to 23:00-7:00 from Monday to Saturday, and to the whole Sunday as well as holiday days. Analysed data lie in the period between 2010 and 2016 for a total of over 21.000 values.

Functions clusterization hypothesis

In this work, a quite general clusterization hypothesis has been made due to the heterogeneity of the building stock. As a first attempt, the entire stock has been subdivided into 9 groups depending on the functions of the analysed buildings. The 9 chosen types are: Agrarian, Medical and Humanities Departments, Scientific Departments (with laboratories), Scientific Departments (without laboratories), large complexes, administrative offices, libraries and sport infrastructures. The proposed clusters, consequently, have been tested with two types of visualization: the scatter plot and the parallel coordinates. The test consists to separate the chosen cluster from all the other ones in order to define, in a qualitative way, cluster thresholds and to look for anomalies and outliers.

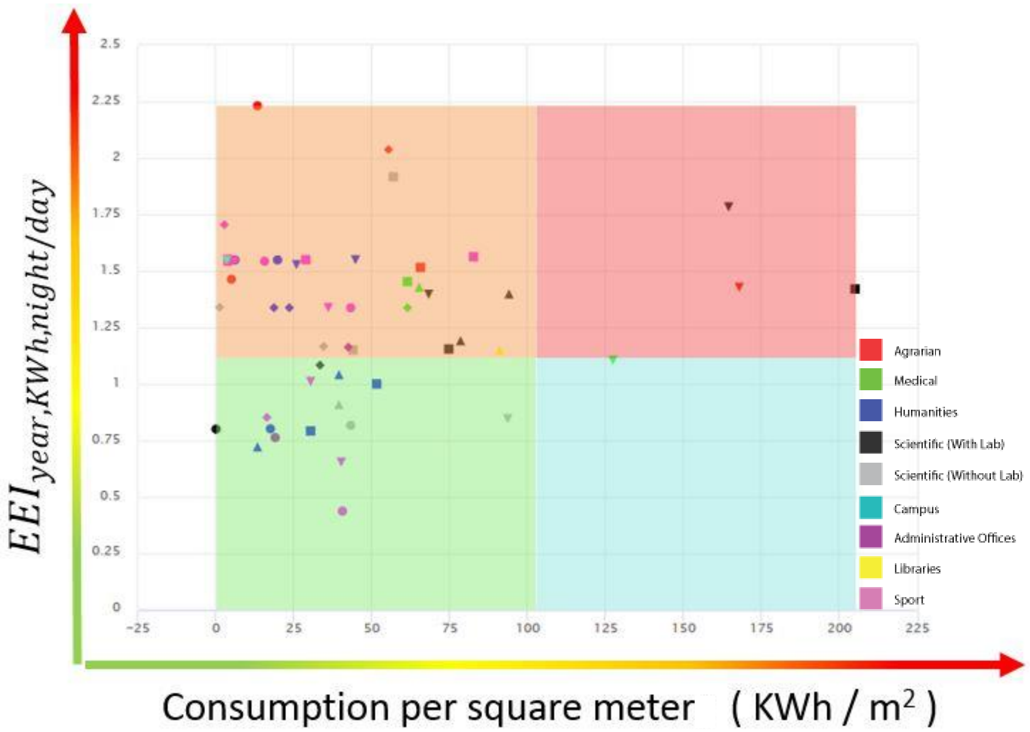

Figure 4. Scatter method graph. X-axis represents consumption per square meter while y-axis corresponds to the night/day energy efficiency index. The four coloured quadrants show four intervention priorities. Nine functions are plotted. Each building type has been shown with a single colour: agrarian (red), medical (green), humanities (blue), scientific with laboratories (black), scientific without laboratories (gray), large complexes (yellow), administrative offices (purple), libraries (aqua) and sport (violet).

Tested methods

Scatter plot approach. The scatter plot has been tested with two pairs of variables for x and y-axis. The first chosen pair, as described by Corgnati et al. [24] is $KWh\ vs\ KWh/m^2$. With this particular choice, as shown in Fig. 1., the graph can be split into 4 quadrants identifying four intervention priorities. The separation lines among the four quadrants correspond to the half of the maximum values for x or y. Unfortunately, with this pair choice there is no possibility to identify clear clusters. This conclusion is also evident by observing the colours: since each colour stands for a different function and all colours are spread all over the plot, the clusters overlap. Going beyond this approach, a second pair has been tested, comparing the energy consumption per square meter and the “night/day energy efficiency index” $KWh/m^2\ vs\ EEI_{year,KWh,night/day}$. In this case, as shown in Fig. 4, the result completely changes with respect to Fig. 1. At a first sight, the points are no longer concentrated in the low priority quadrant but the clusters still overlap.

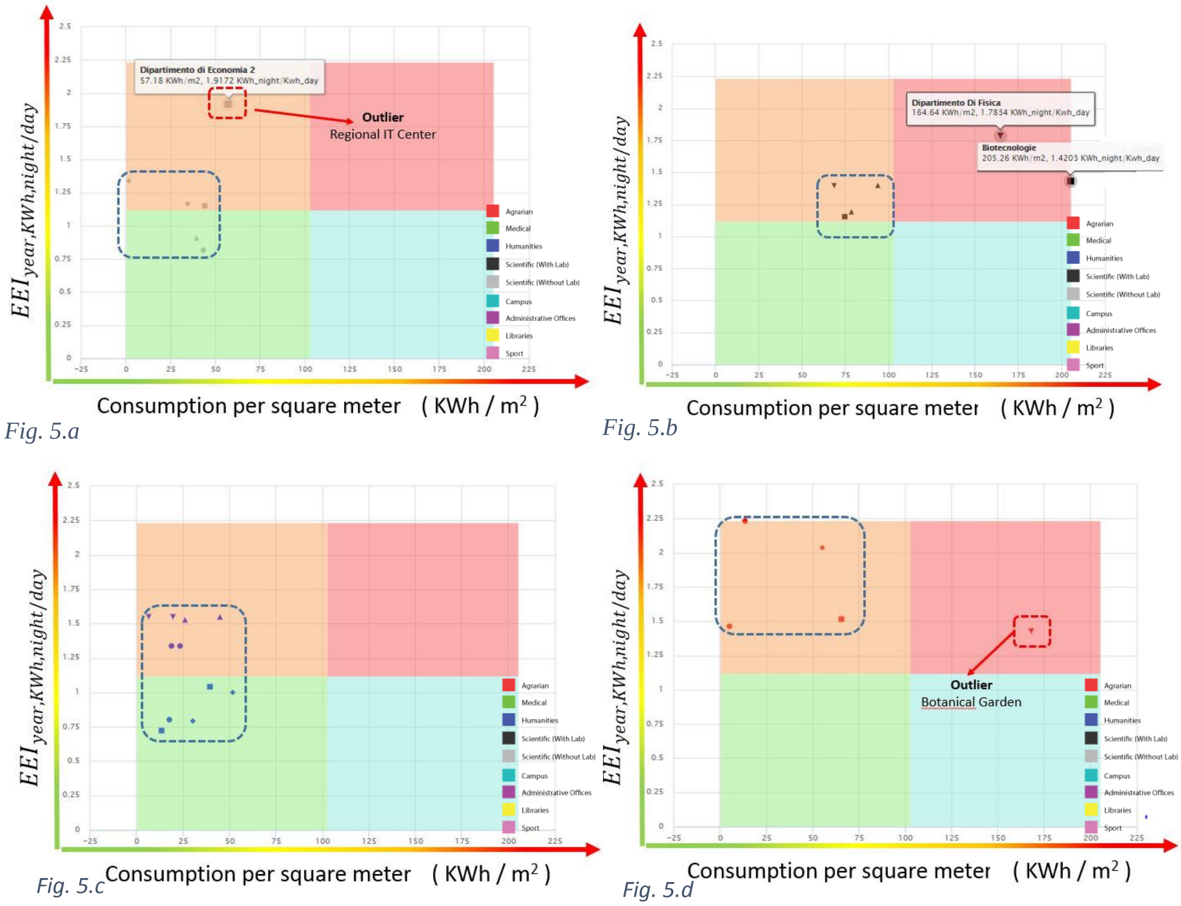

Figure 5. Examples of clusters with different functions. Fig. 5.a (top-left) shows “scientific buildings without laboratories”, Fig. 5.b (top-right) “scientific buildings with laboratories”, Fig. 5.c (bottom-left) “humanities buildings” and Fig. 5.d (bottom-right) “agrarian buildings”.

In order to overcome the overlapping of the clusters, Fig. 5.a-b-c-d show the same graph with four single separate functions. Fig. 5.a-b-c-d, show the “scientific departments (without laboratories)”, “scientific departments (with laboratories)”, “humanities departments” and “agrarian departments”, repsectively. Clusters can be identified quite clearly and our hypothesis can be verified; moreover, energy consumption thresholds can be defined and some outliers can be quickly identified. Roughly speaking, observing the graph, some raw thresholds may be outlined. Fig. 5.a, in fact, shows that “scientific departments (without lab)” have LOW consumption at night and LOW consumption per square meter. Fig. 5.b. exhibits that “scientific departments (with lab)” have MEDIUM consumption at night and MEDIUM consumption per square meter, while Fig. 5.c. points out that “humanities departments” overlap to “scientific departments (without lab)” with a wider range on night/day energy efficiency index. Finally, Fig. 5.d reveals that “agrarian buildings” have HIGH consumption at night and LOW consumption per square meter. The definition of low and high, as explained in previous sections, derives from the four priority zones. In particular low consumption at night corresponds to a $EEI_{year,KWh,night/day} < 1.1$, while low consumption per square meter corresponds to a value $< 100\,KWh/m^2$. On the contrary, high consumption at night corresponds to a $EEI_{year,KWh,night/day} > 1.1$, while low consumption per square meter corresponds to a value $> 100\,KWh/m^2$. On the contrary, high consumption at night corresponds to a $EEI_{year,KWh,night/day} > 1.1$, while low consumption per square meter corresponds to a value $> 100\,KWh/m^2$. The rough medium definition has been given to clusters lying in two different priority zones, or lying close to the limit of a priority zone. More precisely, thresholds can be set for energy consumption per square meter and for day/night energy efficiency index according to Table 1.

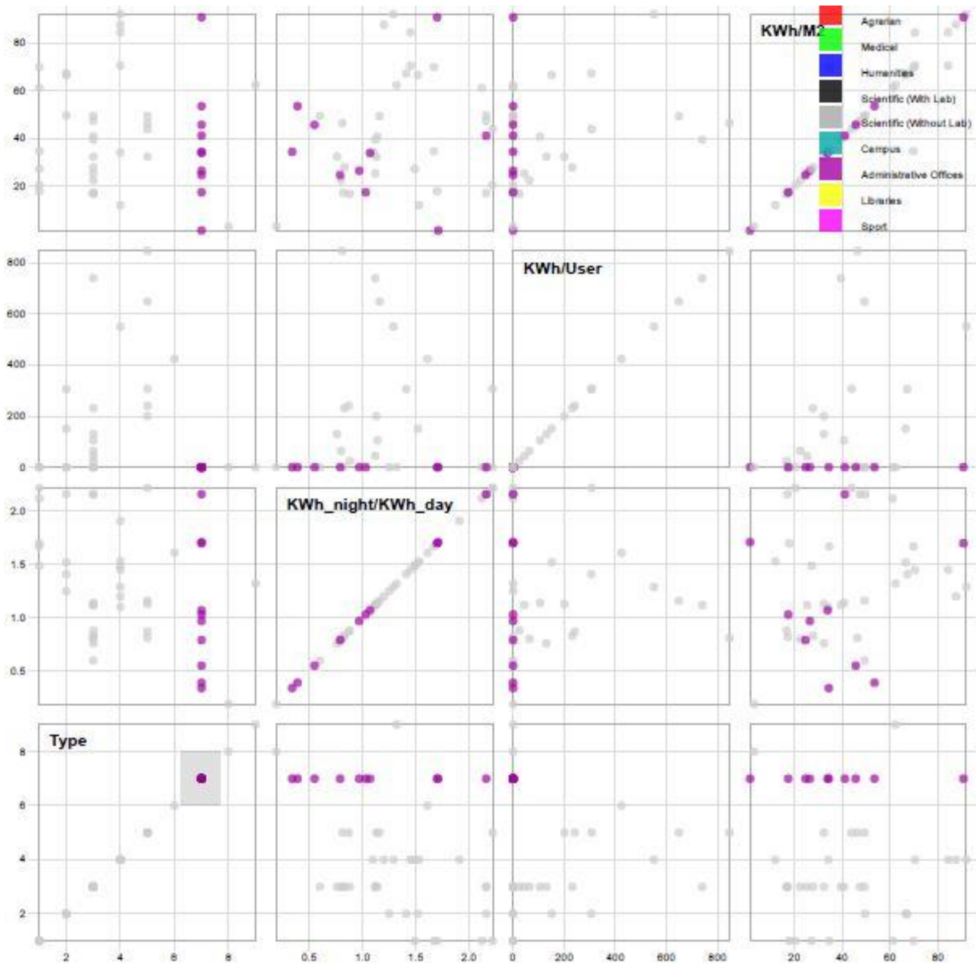

Figure 6. Scatter plot matrix approach. In the figure above the same dataset based on University of Turin building stock has been represented. For each figure, four variables have been shown: type of building (1-9), the night/day energy efficiency index, the electrical energy consumption per user and the electrical energy consumption per square meter. A focus on a building type has been highlighted, thanks to brushing function. The administrative offices have been plotted in violet.

Finally, as an illustration, a few outliers can be quickly identified. Fig. 5.a. highlights an outlier which can be simply explained. In fact, the point represents the Regional IT Center (the electric power is in common with some university classrooms); thus, it cannot be clusterized within the "scientific departments (without lab)" group. Fig. 5.b manifests two outliers, the physics department and the biotechnology department, both with a large IT center and electric chillers. Finally, Fig. 5.d points out the botanical garden as an outlier, with an excessive electric power consumption per square meter. The University of Turin is currently investigating on the reasons of this anomaly. The presented tool is publicly available and it can be consulted online at https://goo.gl/o4nn4f.

Table 1. Thresholds for consumption per square meter and for day/night energy efficiency index.

Building	**Consumption per square meter (KWh / m^2)**	**EEI$_{night/day}$**
Scientific departments (without lab)	< 50	0.75 - 1.3
Scientific departments (with lab)	60 - 100	> 1.1
Humanities departments	< 50	0.7 - 1.6
Agrarian departments	< 75	1.5 - 2.25

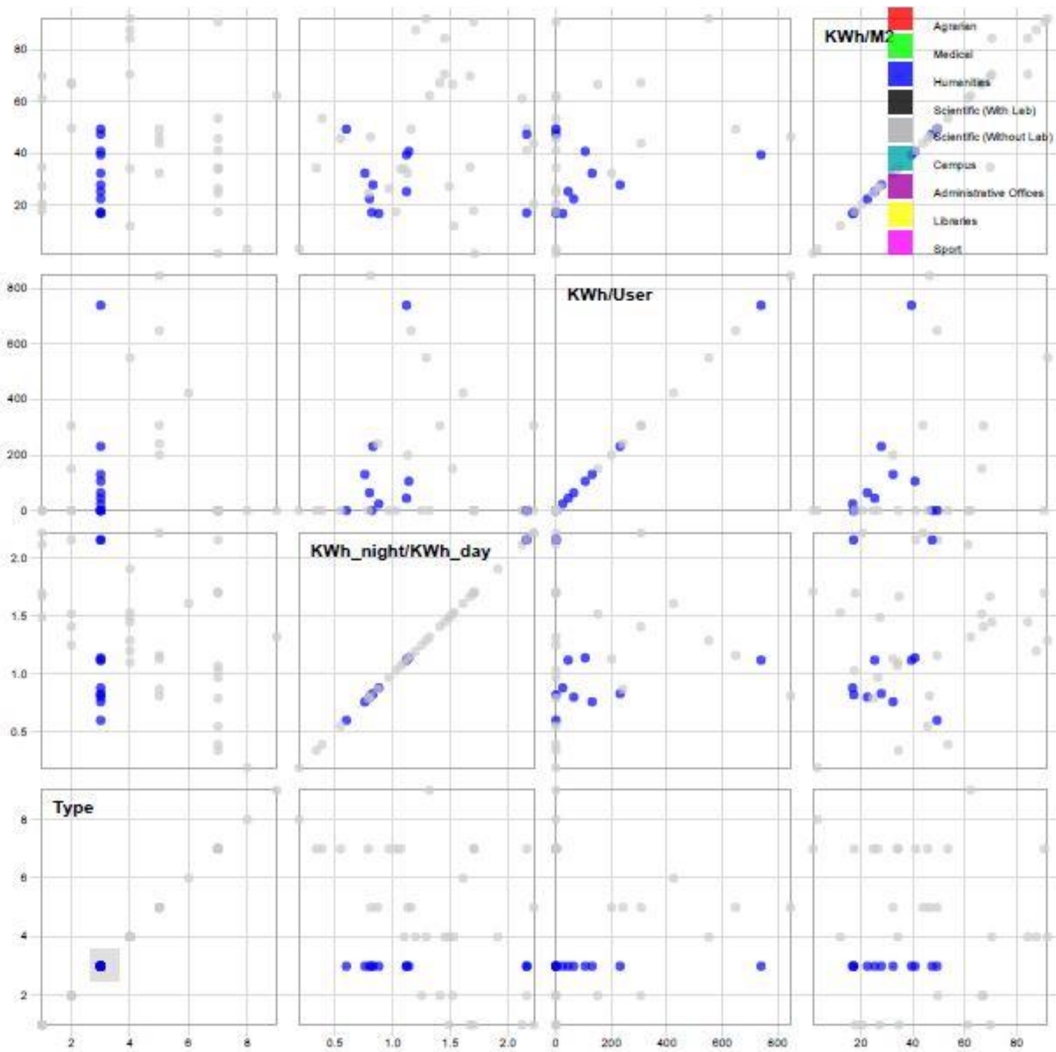

Figure 7. Scatter plot matrix approach. A focus on a building type has been highlighted, thanks to brushing function. The blue dots represent the humanities departments.

Scatter plot matrix. The generalization of the scatter method analysis might be, as previously explained, the scatter plot matrix approach. In Fig. 6 and in Fig. 7, the same dataset, employed in the previous section, is shown as a scatter plot matrix. In this case, more than two pairs of variables can be plotted within the same visualization. As disclosed in Fig. 2, visualizing various variables (dimensions) with hundreds, or thousands, of observations (buildings) can confuse the data miner, but thanks to the brushing function, which allows to select points within a particular range, only a subset of data can be highlighted, reducing a lot the complexity of the visualization. As in a correlation matrix, all plots lying within the diagonal have the same variable for the x-axis and the y-axis. Every box in the left side has the "type of building" variable plotted in the x-axis, whereas graphs in the bottom have the same index in the y-axis. In this case, the data miner can observe the distribution of a precise building type among the different dimensions.

Thanks to this particular visualization configuration, one can check if buildings with the same "building type" belong to a 1-D cluster, or not. This conclusion is possible by observing the point distribution within the left (or bottom) plots. A final observation, for the sake of completeness, has to be highlighted. Scatter plot matrix can be really useful for less than 10 dimensions. In case of more than 10 dimensions have to be analysed, scatter plot matrix suffers of lack of visualization space. The tool here described is publicly available at https://goo.gl/ZJem9h .

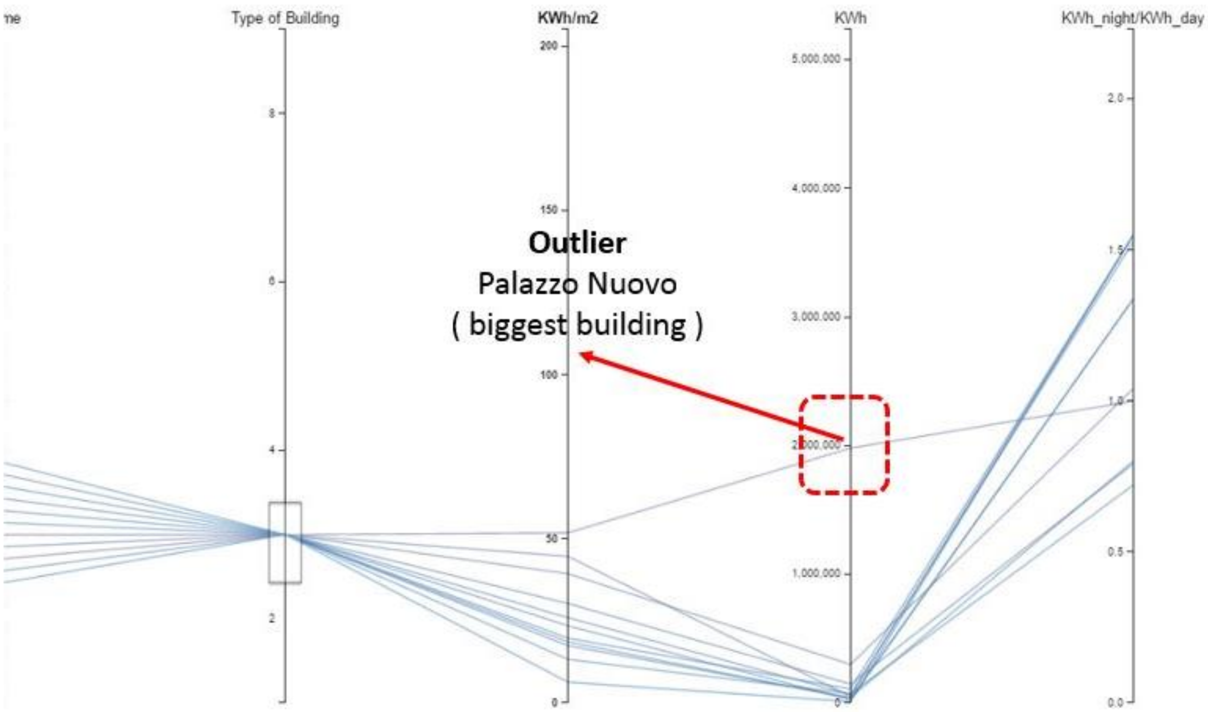

Figure 8. Parallel coordinates method with energy consumption of the UniTo's building stock dataset (Fig. 3). First axis represents building types while the three other axes exhibit annual power consumption, consumption per square meter and night/day energy efficiency index. The gradient of colours is used to visualize different types of buildings. The figure exhibits buildings belong to humanities function, straightforward revealing outliers.

Parallel coordinates approach. The parallel coordinates approach is one of the most powerful tool developed for multidimensional analysis because it allows the simultaneous display of several dimensions and hundreds of observations. This method allows other types of analysis, from historical trend analysis (among different years) to anomalies detection and the simultaneous comparison of variables (dimensions). We consider here two examples, with only few dimensions for visual purposes. The first one is based on the visualization of "Annual electric power consumption", "Consumption per square meter" and the "night/day energy efficiency index". In order to understand the use of the parallel coordinates method, the brush function, one of the most useful features, needs to be discussed. The brush function allows to select a single interval for each dimension.

As described in Data Visualization section, the brushing function allows to exploit the fundamental property "Bumping the boundaries" in order to clearly identify clusters. For this purpose, the first axis, represents building types, from 1 to 9, instead of the different colours described for the scatter method. In this way, it is possible to focus on a particular cluster immediately. Fig. 3 shows the entire dataset analysed. The gradient from red to blue represents, in a graphical way, the 9 building types and it is not correlated to the efficiency of the buildings. At a first sight, one can observe some defined "fluxes" of polygonal lines with a high density. This feature becomes more clear in Fig. 8 and 9 where, respectively, only "humanities departments" and "agrarian departments" have been selected. In these two figures, as in previous methods, the outliers are highlighted in a straightforward way. The parallel coordinates tool here presented is publicly available on: https://goo.gl/4aHYuj .

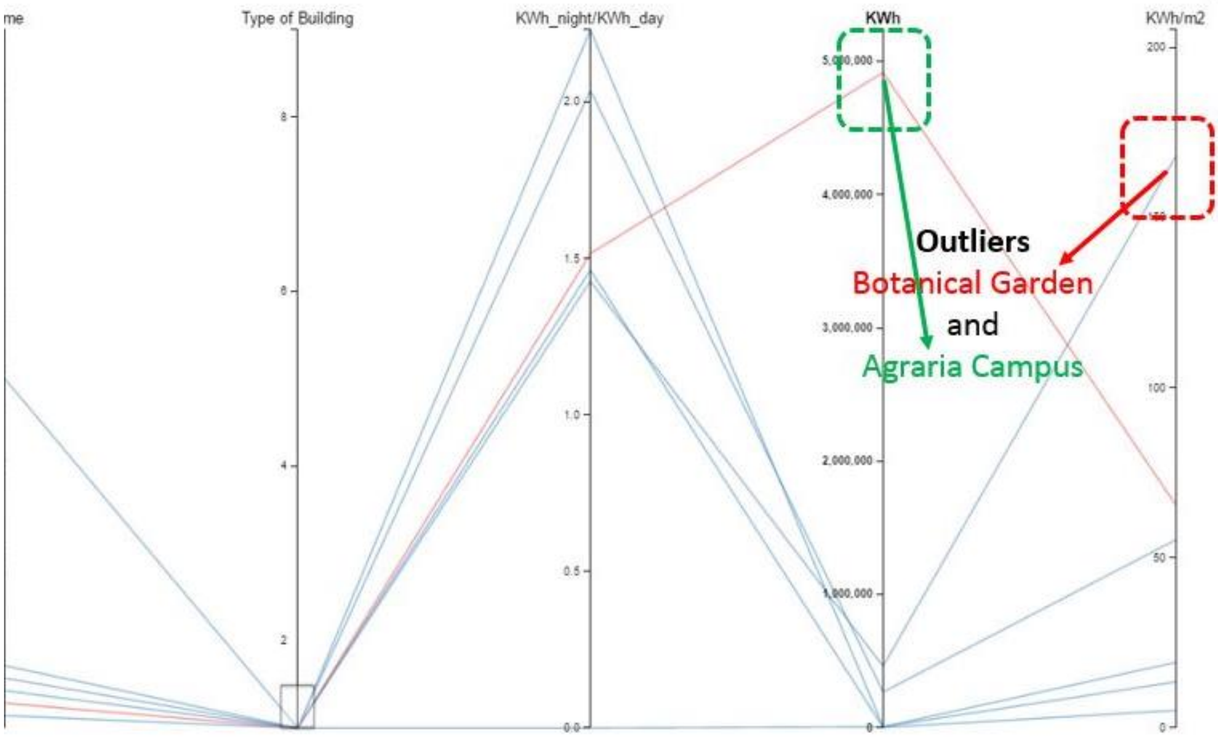

Figure 9. Parallel coordinates method with energy consumption of the UniTo's building stock dataset (Fig. 3). The figure exhibits buildings belong to agrarian function, straightforward revealing outliers.

<u>Historical trends.</u> To conclude, a second application of the parallel coordinates method will be discussed. In particular, in this case, the different axes (dimensions) correspond to different years and only one attribute of the previous case has been plotted. This visualization can be useful in order to track the historical trend of energy consumption, or any other index, in a

unique interactive visualization. Fig. 10 shows the annual electrical power consumption of the whole Unito's building stock. Thanks to this visualization the data miner can observe the historical trend of a particular variable. Moving the mouse over each polyline the real value, and the building ranking for the chosen year is shown.

Moreover, clicking on a particular line, that polyline is highlighted; this feature permits to compare simultaneously the trend of several buildings and to check if some anomalies occur from one year to another (for instance, if the consumption rapidly increases/decreases). Finally, the button, in the bottom-left corner, allows to select and to switch among variables, such as annual power consumption, power consumption per square meter and so on. The tool here described is publicly available at https://goo.gl/YuPTRB.

DISCUSSION

Within previous sections some multidimensional analyses have been presented as useful tools for energy efficiency analyses. Our approach is based on two main considerations and questions. Is it possible to develop quick and user-friendly tools to compare buildings within a large building stock with respect to several energy efficiency indices, trying to simplify the analysis process in order to support energy management of a large public administration? Is it possible to clusterize buildings within a building stock with respect to their specific funtions in order to allow comparison among buildings with completely different architectural features, functions and opening hours? The main aim of this work is to define a general process to qualitatively identify building clusters with respect to energy efficiency indices, exploiting some multidimensional analysis methods. In particular, within this work, we define the thresholds with respect to two main indices, the energy consumption per square meter and the day/night energy efficiency index for four clusters: scientific departments (with and without laboratories), agrarian departments and humanities departments. More precisely two approaches have been exploited, the scatter method and the parallel coordinates.

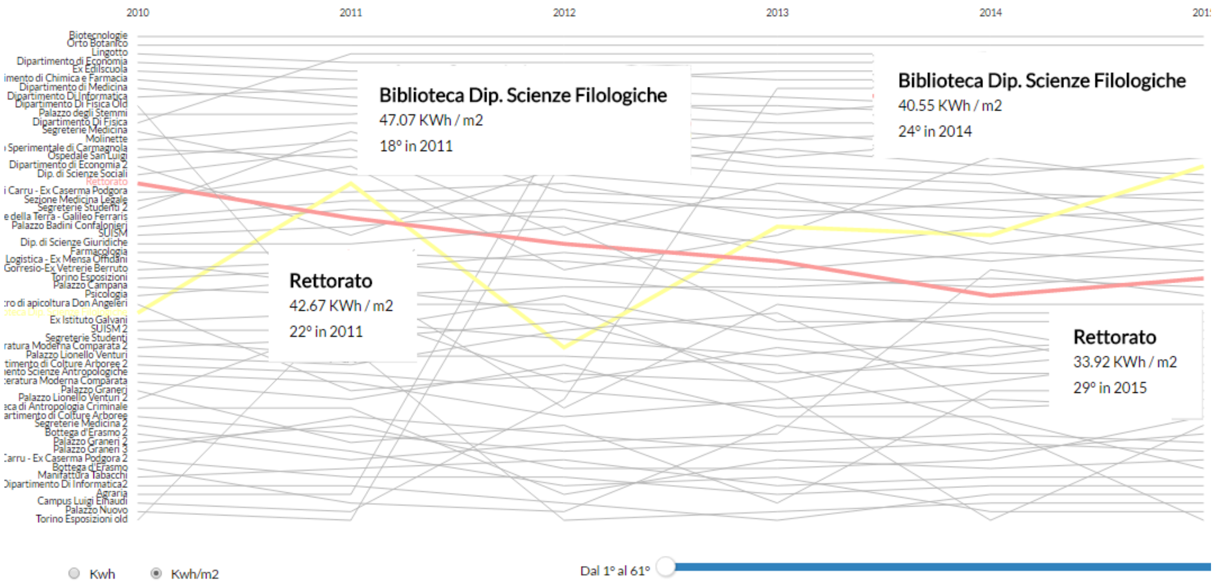

Figure 10. Focused application of the parallel coordinates method on power consumption historical trend. In this particular case each vertical axis (dimension) represents the annual electrical consumption.

The scatter method, described in Data Visualization section, consists in plotting points for each building with respect to two variables (e.g. two indices) on a 2-D graph. Once all points are plotted, the data miner can split the graph into four priority zones and observe the behaviour of each building. Instead, in order to identify clusters, the data miner, or the energy manager, has to assign a building type to each point depending on the main function. In this way, the data miner is able to filter any point with respect to the function and to immediately visualize the cluster, to identify outliers and anomalies. In Fig. 5 four examples relatively to the four cited functions have been presented. As one can notice, the clusters thresholds are quite sharp and outliers can be immediately recognized. The described process needs some tips, recommendations and some background knowledge about the functions. First, the data miner has to know, at least qualitatively, the functions of the buildings. Otherwise clustering algorithms can be used. Second, to exploit the strength of this method, only one function at a time must be visualized. Third, if with the chosen functions there are too many outliers, the cluster hypothesis must be refined because probably the wrong functions have been assigned to the buildings. Finally, once the correct hypothesis is defined, the data miner, or the energy manager, can identify the cluster thresholds for any function. Within Data Visualization section an improvement of the scatter method, called scatter plot matrix has been also described. The scatter plot matrix works in the same way of the scatter method but the data miner can visualize several pairs of dimensions at the same time. The scatter plot matrix exhibits, within all plots in the bottom (or in the left) part of the visualization, the 1-D distribution for each dimension, while all other plots are single scatter plot with a particular pair of dimensions. The major strength of this method is to simultaneously observe the 1-D distribution with respect to several indices for each functions of the buildings. At the same time, the data miner can visualize several scatter plots and try to identify correlations among indices. Finally thanks to the brushing function, described in previous sections, the data miner, or the energy manager, can select and visualize only a subset of the entire dataset. The scatter plot matrix, unfortunately, suffers of a lack of visualization, and it is not possible to visualize more than a dozen of different indices and variables.

Another method, the parallel coordinates, has been discussed. This method has been utilized for two main reasons. The first one, is to overcome the lack of visualization space of the scatter plot matrix. In fact the parallel coordinates method allows to visualize a single attribute on each vertical axis and the method permits to visualize up to dozens of attributes. The second reason is due to the fact to select simultaneously a different range on each axis. In this way we can recognize special features of our dataset, such as simultaneous high-low-high value for different indices or attributes. Fig. 9 illustrates this aspect. The data miner, or the energy manager, can exploit this method to clearly visualize trends and correlations among indices, thanks to the possibility to use the brush function simultaneously on many axes. Finally, the data miner, or the energy manager, may take advantage of the parallel coordinates method in order to visualize the historical trend of a particular chosen index. In this case, the useful feature is to compare simultaneously various buildings with respect to a particular index and observe their historical performance. For example Fig. 10 shows the comparison between two buildings highlighting how the “Rettorato” (the red polyline) is constantly decreasing its consumption per square meter, while the energy consumption per square meter of the “Biblioteca Dip. Scienze Filologiche” (the yellow polyline) is oscillating. Another interesting aspect emerged from this last visualization is to control old and forsaken energy meters. This is possible, visualizing all polylines at the bottom of the graph.

CONCLUSION

Interactive data visualization is a very powerful tool which allows data miners to explore large multidimensional multivariate datasets when no initial hypotheses and no cues are available from previous models and theories. Each tool has to be well-suited and fitted to the specific dataset in a co-design process engaging the final users (for instance, in this case, the energy management). A co-design process may grant to better exploit the potentiality of data visualization, for example avoiding to plot useless variables for a specific purpose. These widely accepted techniques of data visualization may enable energy managers to explore the dataset of a large building stock and to verify their own conclusions in order to better plan possible building structural interventions, or simply to improve the energy management. In this work, a new approach for evaluating the energy efficiency and performance of buildings within a large building stock has been discussed. A very general process has been used to define clusters of buildings depending on their main functions thanks to the scatter plot matrix and the parallel coordinates data visualization methods. Finally, thresholds for the energy consumption per square meter and for the day / night energy efficiency index for any identified cluster have been set. The proposed method aims to simplify the clusterization process with respect to very precise, and time-consuming, GIS analyses; on the other hand, our method aims to design the appropriate method to analyse energy efficiency in a large building stock. As described by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) [27] in their Fundamentals various features must be considered in order to adopt an appropriate method. In particular, ASHRAE suggested to focus on: accuracy, sensitivity, speed, reproducibility, ease of use, level of detail, availability of required data and quality of the output. Specifically, to this recommendation, interactive data visualization may fulfil many of the ASHRAE tips because of high scalability and reproducibility of each designed solution as well as the ease of use, speed and accuracy. Following the ASHRAE recommendations, the process and the approach proposed in this paper are simply based on energy consumption bills, available for any public administration.

NOMENCLATURE

X: entire multidimensional dataset. $X \epsilon R^{mn}$
m: total number of dimensions (attributes) of each observations
n: total number of observations. For this work "observations" represent different buldings.
x_{ij}: real value corresponding to ith attributes of the jth building.
X_j: jth axis in parallel coordinates
O_k: k-th cluster (subset) of X
EEI: generic energy efficiency index
E: generic energy consumption related factor
Λ: generic factor related to energy
E_{KWh}: electrical energy consumption
E_{gas}: gas energy consumption
Λ_{m^2}: building surface
Λ_{users}: number of user for a building
Λ_{DDx}: degree days
$EEI_{year,KWh,\frac{night}{day}}$: annual electrical night/day energy efficiency index
N: number of months
$EEI_{i,KWh,night/day}$: electrical night/day energy efficiency index related to month i.
$E_{i,KWh,night}$: electrical energy consumption at night / weekend / holiday related to month i
$E_{i,KWh,day}$: electrical energy consumption during working hours related to month i

REFERENCES

1. Powell, J. B., Green Building Services., *Journal of International Commerce and Economics.* Published electronically October 2015
2. Balaras, C., Dascalaki, E., Gaglia, A., Droutsa, K. and Kontoyiannidis, S., Energy performance of European buildings, *Proceedings of the Energy Sustainability Conference*, Long beach, USA, pp. 387-396.
3. Poel, B., van Cruchten, G. and Balaras, C.A., Energy performance assessment of existing dwellings, *Energy and Buildings* 39, 393-403, 2007
4. U.S. Energy Information Administration, *Annual Energy Review 2010: October 2011*, Tech. Rep. DOE/EIA-0384, 2011
5. Yun, G.Y., and Steemers, K., Behavioural, physical and socio economic factors in household cooling energy consumption, *Applied Energy*, vol. 88, no. 6, pp 2191-2200, 2011
6. Harris, Z., Higher Education Strategic Planning for Sustainable Development: A Global Perspective, *Springer International Publishing, Handbook of Theory and Practice of Sustainable Development in Higher Education, World Sustainability Series*, DOI 10.1007/978-3-319-47877-7_11, 2017
7. Bakar N.N.A., Hassan M. Y., Abdullah H., Rahman H.A., Pauzi M. D., Hussin F., et al. Sustainable energy management practices and its effect on EEI: a study on university buildings, *Proceedings of global engineering , science and technology conference,* Dubai: UAE, isbn: 978-1-922069-21-4, 1–2 April, 2013.
8. Moghimi S. F. A., Mat S., Lim C.H., Salleh E., Sopian K., Building energy index and end-use energy analysis in large-scale hospitals case study in Malaysia, *Energy Effic*, 7:243–56, 2014.
9. Gonzalez A. B. R., Diaz J. J. V., Caamano A. J., Wilby M. R.,Towards a universal energy efficiency index for buildings, *Energy Build*, 43(4): 980–7, 2011.
10. Li-Ming W., Bai-Sheng C., Modeling of energy efficiency indicator for semi-conductor industry. *Proceedings of the IEEE international conference on industrial engineering and engineering management*, 2007.
11. Cottafava, D., Gambino, P., Baricco, M. and Tartaglino, A., Energy efficiency in a large university, *Sustainable Built Environment. Towards Post Carbon Cities*, ISSN 2283-8791, pp 92-101, July 2016
12. Sonetti G., Lombardi P., Chelleri L., True Green and Sustainable University Campuses? Toward a Clusters Approach, *Sustainability*, 8, 83; doi:10.3390/su8010083, 2016
13. National Bureau of Statistics of China (NBSC), China Statistical Yearbook, *China Statistics Press: Beijing*, China, 2012
14. Mutani G., Delmastro C., Gargiulo M., Corgnati S.P., Characterization of building thermal energy consumption at the urban scale, *71st Conference of the Italian Thermal Machines Engineering Association*, ATI2016, 14-16, Turin, Italy, 2016
15. Inselberg, A., Multidimensional Detective, *IEEEE*, 1997
16. Card, S. K., Mackinlay, J., and Shneiderman, B., Readings in Information Visualization: Using Vision to Think. *Morgan Kaufman*, 1999
17. Keim., D.A., Visual techniques for exploring databases, I*nvited tutorial, Int Conference on knowledge discovery in databases*, KDD97, 1997.
18. NIST/SEMATECH, E-Handbook of Statistical Methods, *http://www.itl.nist.gov/div898/handbook*, 2003
19. Inselberg, A., The plane with parallel coordinates. *Visual Computer*, 1:69-97, 1985
20. Feiner, S., and Beshers, C., Worlds within worlds: metaphors for exploring n-dimensional virtual worlds, *In Proceedins of the 3rd annual ACM SIGGRAPH symposium on User interface software and technology*, pages 76-83, 1990
21. Cleveland, W.S., Visualizing Data. *Hobart Press*, New Jersey, 1993

22. Borg, I., and Groenen., P. J. F., Modern Multidimensional scaling: theory and Applications (second edition). *Springer*, New York, 2005.
23. Bakar, N. N. A.,Yusri, H. M., Hayati, A., Rahman, H. A., et al, Energy efficiency index as an indicator for measuring building energy performance: A review, *Renewable and Sustainable Energy Reviews*, 44, 1-11, 2015
24. Ariaudo F., Balsamelli L., Corgnati S. P., Il Catasto Energetico dei Consumi come strumento di analisi e programmazione degli interventi per il miglioramento dell’efficienza energetica di ampi patrimoni edilizi, *48° Convergno Internazionale AICARR*, Baveno, 22-23 settembre, pp. 547-559, 2011
25. Keller, P. R., Keller, M. M., Visual Cues-Practical Data Visualization. *IEEE Computer Soc. Press*, Silver Spring, 1993
26. Inselberg, A., The plane with parallel coordinates. *Visual Computer*, 1:69-97, 1985
27. ASHRAE, Fundamentals American Society of Heating, Ventilating and Air- Conditioning Engineers. Atlanta, GA, 1997.