Article

EXPLORATIVE MULTIDIMENSIONAL ANALYSIS FOR ENERGY EFFICIENCY: DATAVIZ VERSUS CLUSTERING ALGORITHMS

Dario Cottafava [1], Giulia Sonetti [2], Paolo Gambino [3] and Andrea Tartaglino [4]

[1] Department of Culture, Politics and Society, University of Turin; dario.cottafava@unito.it
[2] Interuniversity Department of Regional & Urban Studies and Planning, Politechnic of Turin; giulia.sonetti@polito.it
[3] Department of Physics, University of Turin; paolo.gambino@unito.it
[4] Energy Management, University of Turin; andrea.tartaglino@unito.it

Abstract: In this paper, we propose a simple tool to help the energy management of a large buildings stock defining clusters of buildings with the same function, setting alert thresholds for each cluster, and easily recognizing outliers. The objective is to enable a building management system to be used for detection of abnormal energy use. First, we framed the issue of energy performance indicators, and how they feed into data visualization (Data Viz) tools for a large building stock, especially for university campuses. Both for Data Viz and clustering algorithm processes, we discussed two possible approaches to choose the right number of clusters and the identification of alert thresholds and outliers, after a brief presentation of the University of Turin's building stock case study. Different Data Viz tools have been studied to apply a specific clustering algorithm, the k-means one. An explorative analysis based on the general Multidimensional detective approach by Inselberg has been performed. Two multidimensional analysis tools, the Scatter Plot Matrix and the Parallel coordinates method have been used. Secondly, the k-means clustering algorithm has been applied on the same dataset in order to test the hypothesis made during the explorative analysis. Data Viz techniques developed in this study revealed to be very useful to explore quickly and simply a large buildings' stock, identifying the worst efficient buildings and clustering them according to their distinct functions.

Keywords: Energy Efficiency Indices, Data Visualization, Clustering Algorithms, University Campus, Energy Management

1. Introduction

Energy efficiency programs as well as policies for the reduction of greenhouse gas (GHG) emissions have been worldwide adopted by national and international governments and public administrations [1]. Reduction of energy consumption and the shift toward a more sustainable use of resources are increasingly becoming a challenge for any sector and activity related to the built environment [2].

The buildings sector is indeed a high energy-consumer, accounting for over one-third of the global final energy consumption [3]. Energy demand is expected to rise by 50% by 2050 if no action is urgently taken [4]. This means that major efforts are required to go beyond existing technical and economic barriers for improving the efficiency of our energy use in buildings. The power to characterise the energy consumption of a complex building stock, for instance, can reduce cost barriers for energy efficient solutions. The improvement of reliable indicators to measure building energy performance at

a neighbourhood/city scale is therefore an important contributions for achieving urban sustainability targets [5,6].

Abu Bakar et al. [7] proposed to measure buildings' energy performance basing on heating, ventilating and conditioning (HVAC) system consumption. Moghimi et al. [8] studied commercial buildings, analysing indicators related to the occupied air conditioning area. González et al. [9] suggested to adopt a reference building in order to compare the energy consumption within a buildings stock.

Although Energy Efficiency Indices (EEIs) are widely studied, there is still a lack of researches in literature related to energy decision-making tools relying on these indices [10]. For this reason, current research challenges are envisaged in developing links between EEIs and more general energy assessment frameworks, to enable sounding comparisons among buildings with different architectural features, functions and/or occupations schedules [11].

To this respect, university campuses may represent a valuable test bed, being often a joint resemble of building with very different characteristic yet belonging to a same purpose. For their physical scale in the city, university's campuses have a significant role to play respect to local energetic and socioeconomic impacts, going far beyond the university scale itself [12]. Universities are increasingly conceived as hubs for innovation, serving as test bed for new energy reduction strategies [13–15].

However, a mayor focus among all the initiatives is generally devoted to energy performance improvement, and its monitoring [16], justified by the increased investments in energy efficient technologies [17]. Living labs monitoring infrastructure provide an appropriate way for answering energy data queries while displaying all the necessary information for performances self-assessment and external reporting purposes [18]. There is, however, a gap between these energy performances oriented experiences and the international ranking systems for green labelling of campuses which are not based on performance indicators but relying on ranges of total energy consumption [19].

Towards the same direction, a work of the National Bureau of Stastistics of China [20] highlights that universities or megaversities with different building functions have energy consumption per square meter that cannot be compared and classified with the same criteria. Those challenges are also linked to the diversity of material utilization, CO2 emissions, energy source and regulatory compliance, which is different from country to country, and from city to city [21].

1.1. Motivation and problem identification

Both at city scale or campus scale, as already noted by Haas [22], the most difficult task when dealing with EEIs is to provide the corresponding data by end use to obtain suitable numbers for cross-country evaluations. Many of the parameters needed for time series and cross-country analyses are dependent on the obtainability of disaggregated data from wide-ranging surveys and cross-section analyses, and there are number of critical methodological problems that still pave the way for the creation of such operational indicators of energy efficiency [23]. Regarding the specific University campus realm, Sonetti et al. [24] already argued the lack of a precise analysis based on building types or functions, in one of the most spread and recognized green ranking for universities, the *UI GreenMetric - World Universities Rankings*. The need of three clusterizations based on urban morphology, climate zones or university functions has been highlighted for a sounding comparisons among campuses.

However, the advantages of performing large scale energy monitoring through easy visualization tools are many, for example, the association of a fixed (or predictable at least) amount of energy resources in areas of a city or in different buildings of the same district [25], energy outliers [26],demand side management operations and local balancing [27], entrants for critical retrofit intervention [28], large benchmarking analysis engaging allowing inter-comparison [29,30], peak power demand [31], and so forth.

1.2. Current paper aim and structure

The paper's aims are two-folds: propose a simple, efficient and precise analysis tool able to compare buildings within a large stock, inputting only energy efficiency indices; and explore how to use this tool to cluster buildings within a stock according to their specific function. The proposed tool tries to fill the gap between very detailed energy audits analysis and the lack of precise user-friendly and immediate tools for energy efficiency comparisons among buildings. The proposed approach needs basic energy data input for each building - i.e. monthly energy bills – and, starting from those, it adopts interactive data visualization tools to analyse the dataset. The *multidimensional detective* approach, as described by Inselberg [32], has been adopted to define the clusters' alert thresholds.

The paper is structured as follows. First, we framed the issue of energy performance indicators, and how they feed into data visualization tools for a large building stock, especially for university campuses (Introduction). In the Large scale buildings energy monitoring methods section, current Data Visualization techniques and clustering algorithms are explained. In the Methodology section, the adopted approach for developing a simple energy monitoring tool exploiting the University of Turin's building stock, defining clusters of buildings with the same function, setting alert thresholds for each cluster, and easily recognizing outliers is described. Both for data visualization and the clustering algorithm processes, we discussed two possible approaches to choose the right number of clusters and the identification of alert thresholds and outliers, after a brief presentation of the University of Turin's building stock case study. Finally, Results and Discussion report a comparison between the two approaches with considerations on the obtained clusters and their accuracy.

2. Large scale buildings energy monitoring methods

2.1. Data Visualization

In the Big Data decade, data visualization becomes fundamental to extract useful and valuable information from the enormous amount of data available today. Each specific dataset, in fact, potentially has a huge amount of hidden information and could reveal important tips for managers and policy makers, as well as for data miners and data scientists. According to Card et al. [33], *Information Visualization*, the most general definition of Data Visualization (DataViz), is defined as visual representations, computer-supported, able to amplify human cognition. Keim et al. [34], in fact, define DataViz as the process to "translate" complex dataset into visual tips and immediate qualitative information and they identify three main aims: presentation, confirmative and explorative. For both three aims, one of the fundamental aspects of DataViz is based on the interactive process allowed by modern DataViz coding libraries, as D3.js [35], Julia [36], GoogleCharts and others tools, which permit users to manipulate datasets in order to better understand hidden information in datasets. Within this framework, interactive Data Visualizations are crucial for explorative analysis where data miners have no quantitative insights to model a particular datasets. This is particularly important for data driven researches as for energy efficiency studies, or more in general for analysis aimed at policy makers and managers, where the main aim of an analysis should be to identify alert thresholds, outliers or anomalies [37].

Generally speaking, each multidimensional dataset X is composed by n arrays - i.e. the number of observations/the size of the dataset, $x_i = (x_{i1}, x_{i2}, ..., x_{im}), i = 1, ..., n$ with m attributes/dimensions and it may be represented by a matrix nxm. With this representations x_{ij} is the datum of the real observation i with attribute j. Data Visualization techniques may be grouped into four main approaches: 1) Axis reconfiguration [38], 2) dimensional embedding [39], 3) dimensional sub-setting [40] and 4) dimensional reduction [41]. In particular, two approaches out of four - i.e. axis reconfiguration and dimensional sub-setting - will be discussed within this paper, exploiting respectively the *Scatter Plot Matrix* (dimensional sub-setting) and the *Parallel Coordinates* (axis reconfiguration), two of the most popular techniques.

The Scatter Plot Matrix

It highlights, as described by Keller [42], relationships among variables as in a correlation matrix, where single scatter plots between two attributes of the datasets are plotted within the same graph. The Scatter Plot Matrix can be understood as a generalization of a single Scatter Plot. With respect to the energy field, for instance, Corgnati et al. [43] proposed the use of a single Scatter Plot based on two attributes - i.e. the annual building consumption and the annual electrical building consumption per square meter - in order to identify the top interventions priorities within a large building stock, while Cottafava et al. [44] proposed two other attributes in order to identify buildings with the most inefficient lighting and heating schedules: electrical building consumption per square meter and the day/night energy efficiency index (a ratio between energy consumption during the weekday working hours and during the night/weekend). Thus, the Scatter Plot Matrix could be exploited as a preliminary analysis method useful to identify the top/bottom priorities with respect to three, or more, attributes of a datasets.

The Parallel Coordinates

This method, introduced by Inselberg [38], allows to visualize a muldimensional dataset thanks to m equidistant copies of the y-axis, perpendicular to the x-axis. Thanks to this method, the observation $x_i = (x_{i1}, x_{i2}, ..., x_{im})$ is represented as a polygonal line which intersects each vertical axis. It is noteworthy to highlight that, in this visualization, each vertical axis represents a different attribute/dimension of a multidimensional dataset, and each polyline represents a different observation. In order to exploit the Parallel Coordinates method is crucial to cite one fundamental property, named *Bumping the Boundaries*, which ensures that a polygonal line lying in-between two other polygonal lines, it represents an interior point of the corresponding hypersurface in m dimensions [32].

2.2. Data Clustering algorithms

Data Clustering is a process of detection of different groups within a specific dataset in order to identify patterns or subsets, i.e. clusters, as well as outliers. Clustering process aims to identify clusters where *"Instances, in the same clusters, must be similar as much as possible"*, meanwhile *"Instances, in different clusters, must be different as much as possible"* [45]. Clustering, in particular, is an unsupervised process where instances (objects) have no initial label (i.e. assigned cluster) given by data scientists and researchers but the cluster configuration depends on the chosen algorithm and on the adopted *similarity measures* and *distance metrics*.

Distance metrics

Metrics depend on, as reviewed by Xu et al. [46], the adopted definition of distance. The most common used definition, for quantitative measures, is the *Minkowski distance* of order p:

$$D_{ij} = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^{\frac{1}{p}} \right)^{p}$$

where $d = n.\ of\ dimensions$, $x_{ij} = value\ of\ the\ attribute\ j\ of\ the\ object/point\ i$ and D_{ij} is the distance between the point i and the point j. For precise p the Minkowski distance is defined as the *Euclidean distance* (Minkowski order 2), the *Manhattan distance* (order 1) or the *Cebysev distance* (order ∞). Other common distance metrics are based on the *Mahalanobis distance*, $D_{ij} = \left(x_i - x_j\right)^T S^{-1} \left(x_i - x_j\right)$ and the *Jaccard distance* $J_\delta\left(A, B\right) = 1 - |A \cap B|/|A \cup B| = |A \cup B| - |A \cap B|/|A \cup B|$ where S is the Covariance Matrix of the cluster where x_i and x_j belong to the same group and $|X|$ is the number of element in subset X [47].

Evaluation

Evaluation consists in the process of testing of the validity of the chosen algorithm. Evaluation indicators may be subdivided into two categories: *internal evaluation* and *external evualuation*. The first one refers to data within the same cluster, while the second one refers to similarity evaluation among data lying in different clusters [47]. Some of the most widely adopted internal evaluation methods are: i) the *Within-Cluster Sum of Square* [48]

$$Q_T = \frac{1}{k}\sum_{j=1}^{k}\sigma_j = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{|Z_j|}\frac{d(x_i^j, c_j)}{|Z_j|} \tag{1}$$

ii) the *Davies-Bouldin Index* [49]

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d\left(c_i, c_j\right)}\right) \tag{2}$$

iii) the *Silhouette Index* [50]

$$S = \frac{1}{k}\sum_{j=1}^{k}S_j = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{|Z_j|}\sum_{i=1}^{|Z_j|}\frac{b_i^j - a_i^j}{\max\left[a_i^j, b_i^j\right]} \tag{3}$$

where

$$a_i^j = \frac{1}{|Z_j|}\sum_{l=1, l\neq i}^{|Z_j|} d\left(x_i, x_l\right) \text{ and } b_i^j = \min_{p=1,\ldots,k; k\neq j}\left[\frac{1}{|Z_p|}\sum_{l=1}^{|Z_p|} d\left(x_i^j, x_l^p\right)\right]$$

where n = *total number of points*, x_i^j = *point i lying in cluster j*, k = *n. of clusters*, c_x = *the centroid of the cluster x*, σ_x = *the mean distance between any data in cluster x and the centroid of the cluster*, $|Z_x|$ = *n. of point in cluster* Z_x, $d\left(x_i, x_j\right)$ = *the distance between points* x_i and x_j (both centroids or observations). Finally, there are *various external evaluation indices*, as reported by Dongkuan et al. [47] (i.e. Rand index [51], Jaccard Index [52], Fowlkes-Mallows Index [53], ...) useful to evaluate the efficiency of clustering algorithms in terms of finding true (false) positives and negatives with respect to a reference cluster configuration.

Clustering Algorithms

In literature, generally, clustering algorithms are mainly split into two main categories - *Hierarchical* and *Partition* clustering methods - but various sub-classifications have been proposed in order to categorize the dozens of clustering algorithms. Dongkuan et al. [47] subdivide algorithms in traditional ones and modern algorithms. Traditional algorithms have been aggregated into 9 categories - partition, hierarchy, Fuzzy Theory, distribution, density, graph theory, grid, fractal and model - based, while for modern algorithms they count more than 40 proposed algorithms divided into 10 categories. Nagpal et al. [54], instead, propose a classification where algorithms are - partition, hierarchy, density, grid, model and category - based. Partition clustering algorithms arrange the n data into k different clusters [55]. The number k of cluster is an input parameter of the algorithm. The partitioning is obtained by minimizing an objective function, and it depends on the distance from the centroid to any point within a single cluster or on some similarity functions. Basically, the initialization of a partition algorithm consists in: a) assigning randomly k seed points, the initial centroids and b) every point in the dataset must be labelled to the nearest cluster centroid. Then, in each step, c) a new centroid for each cluster must be computed by averaging over all points lying in the same cluster and d) the nearest centroid for every point in the dataset must be checked again. Steps c) and d) continue untill a local optimum is found. The two most famous partition clustering algorithms are the k-means [56] and the k-medoids (K-means for discrete data) [57] directly developed from the core concept of partition algorithms. A typical way to choose seed points, for instance, as reviewed by Nagpal et al. [54], is to choose

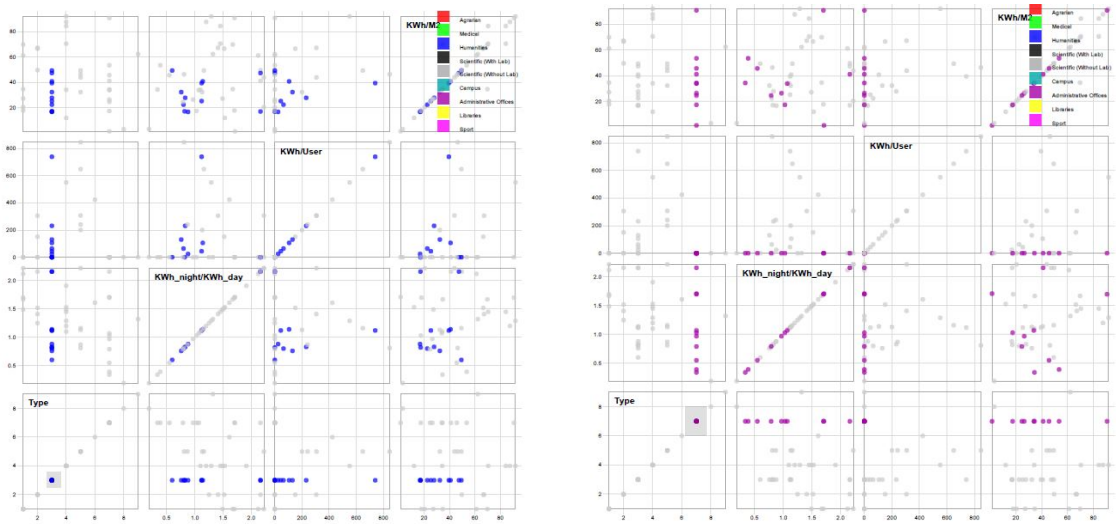
randomly from the existing points, in order to avoid empty clusters. Other partition algorithms, instead, as CLARA [58], CLARANS [59] and PAM [60] choose seed points randomly in a grid based way. Generally, the advantage of these algorithms is a high efficiency and low time complexity while disadvantage consists in the necessity of defining the number of clusters k as an algorithm input, taking into account that the choice of k affects results and the identification of outliers. Hierarchical algorithms find clusters in an iterative way starting from the whole dataset in a unique cluster, *divisive mode* (top-down approach), or from a single point, *agglomerative mode* (bottom-up approach). The basic idea of hierarchical algorithms is to find nested clusters starting from 1 group to n groups or vice versa in an iterative way merging (or splitting) the nearest clusters (or the furthest ones). Typical algorithms are CURE [61], BIRCH [62], CHAMELEON [63] and many others. For instance, BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies - is based on saving only the *Cluster Features* triple n, LS, SS where n =*total number of points within a cluster*, LS is the sum of attributes of all points within a cluster and SS is the sum of square. CURE - Clustering Using REpresentatives - thought for large database, is insensitive to outliers, while CHAMELEON merges two cluster only if they are close "enough". Many algorithms, such as the k-means, need the number of cluster k as an input, while many others determine the right number in a dynamic way. The problem of the identification of the number of clusters can be solved thanks to various methods. For instance, Ketchen et al. [64] analysed the elbow method based on the within-cluster sum of square, method introduced by Robert L. Thorndike [65] in 1953. The elbow method consists in plotting the within-cluster sum of square, i.e. the average distance of any point within a cluster with respect to its centroid, in a scatter plot with the number of cluster k, looking for the "elbow", the point where the WSS stops to rapidly decrease. The elbow point shows the best number of cluster k. Pollard et al. [66] use the Mean Split Silhouette (MSS), a measure of cluster heterogeneity, and they minimize it to choose the best k. Tibshirani et al. [67], instead, proposed the gap statistic, a methodology based on the comparison of the change in within-cluster sum of square dispersion with respect to a proper reference null distribution. Other methods, widely adopted in literature, are based on MonteCarlo simulations cross validation [68,69]. Consensus Clustering [70] and Resampling [71] try to find k looking for the most "stable" configuration through different MonteCarlo simulations but with the same number of clusters. On the contrary, Junhui Wang [72] proposed to select the number of clusters minimizing algorithm's instability, a simple measure of the robustness of any algorithm against the initial random seeds.

3. Methodology

In order to design a simple, user-friendly approach for energy efficiency analysis for large buildings stock, we compared different Data Visualization tools applying a specific clustering algorithm, the k-means one. An explorative analysis based on the general *Multidimensional detective approach* [38], has been performed as first step. We exploited two multidimensional analysis tools, the Scatter Plot Matrix and the Parallel coordinates method. Secondly, the k-means clustering algorithm has been applied on the same dataset in order to test the hypothesis made during the explorative analysis. The first step, the multidimensional detective approach as the one proposed by Inselberg [38], identified the most meaningful clusters. As described in Cottafava et al. [44], the process consists of few steps, and it is able to identify outliers and "junk attributes" as well as to define boundaries and alert thresholds, a minimum and a maximum value, such as $x_{min,j} \leq x_{ij} \leq x_{max,j}, \forall x_i \epsilon Z_k$ where Z_k is the k-th subset of X for every cluster. The three steps - i) *define building types*, ii) *test the assumptions* and iii) *identify thresholds and outliers* - consists in choosing the building types (e.g. libraries, hospitals, research centres, and so forth) and labelling each data relying on the knowledge background of the data source organization. When each datum has been labelled, alert thresholds can be identified and outliers can be recognized. The three steps have been accomplished via the Scatter Plot Matrix and the Parallel Coordinates methods. After defining clusters and thresholds, the k-means algorithm tests the validity of the clusters hypothesis. Finally, we propose a tool to monitor historical trends based on an interactive application of the Parallel Coordinates method.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2018 doi:10.20944/preprints201804.0127.v1

Peer-reviewed version available at Energies 2018, 11, 1312; doi:10.3390/en11051312



(a) Humanities Departments **(b)** Administrative Offices

Figure 1. Scatter plot matrix for the Unito's buildings stock with respect to four attributes: type of building (1-9), the night/day energy efficiency index, the energy consumption per user and the energy consumption per square meter.

3.1. Dataset and indices description

As briefly mentioned in the introduction, the selected case study for testing the simple tool for large scale building stocks energy analysis has been the University of Turin (Unito) in Italy. The advantage of choosing the Unito campus relies in the availability of a wide historical data set and the precise match of energy-related information and the locus of its consumption, thanks to a wide net of smart meters, periodical human-based control on data trends and an open access website prompting all data. The University of Turin is a little city within a city: Unito's buildings stock is very heterogeneous with respect to functions of the buildings, their construction year (ranging from the XVI century to 2014) and architectural features. It sums more than $800000 m^2$, with about 120 buildings sprout all over the city and in Piedmont region, for a total of 2.08 TOE of methane gas and 23.5 GWh of electrical energy consumption per year. The buildings stock comprises museums, administrative offices, libraries, hospitals, as well as research centres, a botanical garden and departments of humanities and sciences [73]. The Unito energy data related to a whole year on monthly basis have been adopted as the training dataset for this study. Analysed data refers to 46 buildings, with 59 electricity meters and 77 methane gas meters. Four attributes for each point have been chosen: the absolute annual energy consumption (kWh), the annual energy consumption per meter square ($^{kWh}/_{m^2}$), the annual energy consumption per user ($^{kWh}/_{user}$)and the *"night/day energy efficiency index"* $EEI_{year,kWh,night/day} = {}^1/_{12} \sum_{i=1}^{12} E_{i,kWh,night}/E_{i,kWh,day}$ where $E_{i,kWh,day} = kWh$ *during working hours* and $E_{i,kWh,night} = kWh$ *during night/holiday for month* i.

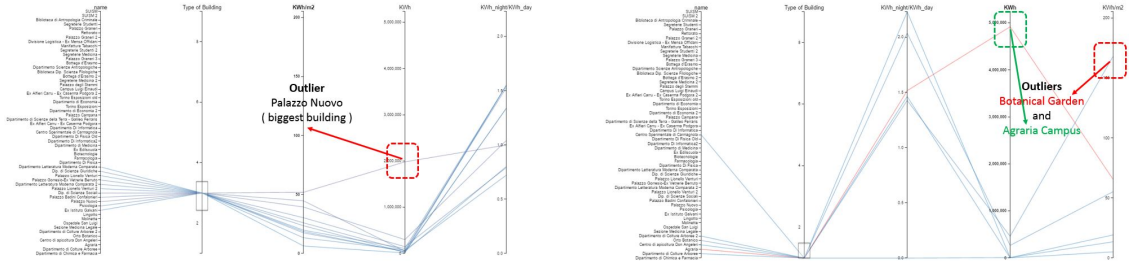
3.2. k-means algorithm

The k-means algorithm has been used for the same dataset in order to compare results obtained by the algorithms with the results obtained by the multidimensional detective approach. Each real observation x_{ij}, for each dimension j has been normalized so that $x_{ij} = (x_{ij} - \min x_j)/(\max x_j - \min x_j) \epsilon (0,1)$, in order to allow to compute a meaningful Euclidean distance metric among points. The initial centroids for each cluster have been picked at random among the existing points of the dataset in order to avoid empty clusters. Three internal evaluation indices have been used to validate results and to choose the right number of clusters k - the within-cluster sum of square, the Davies-Bouldin

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2018 doi:10.20944/preprints201804.0127.v1

Peer-reviewed version available at Energies 2018, 11, 1312; doi:10.3390/en11051312

index and the Silhouette index. The final result, for each k (from $k = 2$ up to $k = 15$), has been chosen as the best configuration - the one with the minimum WSS index - over 1000 independent MonteCarlo simulations. The right number of cluster k, as described by the Elbow method, has been obtained by identifying the *elbow* in the scatter graph *WSS VS k*. Finally, once defined the right k, the best cluster configuration has been selected chosing the highest external evaluation indices, the *Rand Index* and the *Fowlkes-Mallows Index*, over 1000 MonteCarlo simulations, with respect to the algorithm result and the target cluster configuration. The target cluster configuration is the one chosen during the multidimensional detective process.



(a) Humanities Departments **(b)** Agrarian Departments

Figure 2. Parallel coordinates method for the Unito's buildings stock with respect to four attributes: type of building (1-9), the night/day energy efficiency index, absolute annual energy consumption and the energy consumption per square meter.

4. Results

4.1. Cluster Identification

Cluster hypothesis

A general hypothesis has been made due to the heterogeneity of the Unito's building stock. The whole stock has been categorized into nine clusters with respect to the functions of the buildings: Scientific Departments (with laboratories), Scientific Departments (without laboratories), Medical, Agrarian and Humanities Departments, libraries and administrative offices, and, finally, sport infrastructures and large complexes.

Data Visualization Techniques.

The proposed clusters have been tested with two types of visualization: the Scatter Plot Matrix, a dimensional sub-setting method (Fig. 1), and the Parallel Coordinates method, an axis reconfiguration technique (Fig. 2). First, our approach 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. Second, hypothesis have to be tested in order to identify alert thresholds and outliers. The first step can be achieved thanks to the brush functions of the two proposed visualizations. As shown in Fig. 1a and Fig. 1b for the Scatter Plot Matrix and in Fig. 2a and Fig. 2b for the Parallel Coordinates method, the identification of the pre-defined clusters is straightforward and outliers emerge in a very clear way. *The Scatter Plot Matrix* is the generalization of the Scatter Plot, as described in Cottafava et al. [73] and as publicly available at `https://goo.gl/o4nn4f`. Fig. 1 shows the whole buildings' stock of the University of Turin and reports 16 different single Scatter Plots. Respectively x-axis, and y-axis, starting from the bottom-left graph, report the following attributes: *Type of building*, *the day/night energy efficiency index*, *the annual energy consumption per user* and *the annual energy consumption per meter square*. The four graphs on the diagonal, as for a correlation matrix, has the same attribute both on x-axis and

y-axis. Each cluster is identified with a different colour and it can be highlighted simply selecting the type of the building in the bottom-left graph. The nine labelled colours are: red (Agrarian depts.), green (Medical depts.), blue (Humanities depts.), black (Scientific depts. - with lab.), grey (Scientific depts. - without lab.), sky-blue (Large complexes), yellow (Libraries) and pink (Sport infrastructure). In particular, Fig. 1a reports, as an example, the Humanities Departments and Fig. 1b shows the Administrative Offices of the University of Turin. This visualization configuration allows to check if buildings with the same label lye on the same 1-D cluster, simply observing points distribution on the left and bottom plots. The tool here described is publicly available at `https://goo.gl/ZJem9h`. *The Parallel Coordinates* method also allows to display various attributes for hundreds points with a different visualization configuration. This approach permits data miner to analyse dependent, or independent, attributes and to detect anomalies or precise trends and correlation among different attributes as in a pattern recognition problem. Fig. 2 shows the whole Unito's buildings stock with respect to four different attributes: the *type of the building*, the *annual energy consumption per square meter*, the *absolute annual energy consumption* and the *day/night energy efficiency index*. In this case, the nine clusters are labelled with number from 1 to 9 and represented by the first vertical axis. Respectively, from 1 to 9, the clusters correspond to the following: agrarian depts., medical depts., humanities depts., scientific depts. - with lab, scientific depts. - without lab, large complexes, libraries and Sport infrastructure. As for the Scatter Plot Matrix, in this case the brush function allows data miner, or the policy maker/energy manager, to highlight precise subset of the whole dataset. This feature permits to exploit the property *Bumping the boundaries* in order to bound the clusters. Fig. 2a and 2b, respectively, show humanities depts. and agrarian depts. At a first sight, it is possible to notice quite precise fluxes/patterns of polygonal lines with a high density. The tool we used is publicly available on: `https://goo.gl/4aHYuj`.

Clustering Algorithm.

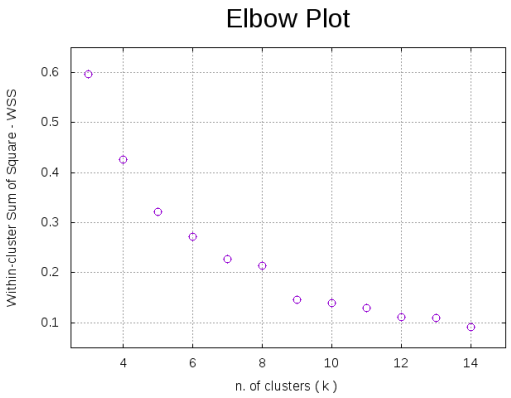


Figure 3. Elbow method. The plot shows within-cluster sum of square VS k (n. of clusters). The right k number is between 9 and 10.

The k-means algorithm has been used in order to identify and recognize clusters depending on three main attributes, *annual absolute energy consumption*, *annual energy consumption per square meter* and the *day/night energy efficiency index*, avoiding the energy consumption per user due to lack of data for administrative offices and other buildings. In this paragraph, first, we report some considerations on the right number of clusters found thanks to the elbow method. We select the best configuration for each k - i.e. the lowest WSS - running one thousand MonteCarlo simulations. The elbow method suggests, as previously defined in data visualization analysis, that the right number of k is between 9 and 10, where the WSS slightly stop to decrease. Fig. 3 shows the elbow plot with the WSS index on the y-axis and k, the number of clusters on the x-axis. In Table 1 we report data obtained related to WSS, to the Davies-Bouldin index and to the Silhouette Index. Silhouette index is slightly constant for different

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 April 2018 doi:10.20944/preprints201804.0127.v1

Peer-reviewed version available at Energies 2018, 11, 1312; doi:10.3390/en11051312

k while WSS and DB index decrease as k increase. Since Silhouette index lies in $-1 \leq Sil \leq 1$, where a Sil index of -1 means a bad cluster correlation and 1 a good one, the obtained clusters represent a quite good configuration.

Table 1. best configuration evaluation index.

k	WSS	DB Index	Sil Index
3	0.597404	2.16486	0.406959
4	0.425536	1.98487	0.504544
5	0.32147	1.96633	0.478912
6	0.271463	1.72633	0.465793
7	0.22802	1.70083	0.48971
8	0.213499	1.69415	0.411259
9	0.146231	1.50094	0.679899
10	0.140161	1.53901	0.531202

Comparison between DataViz and k-means clusters.

Once chosen the best number of clusters ($k = 9$), two external evaluation indices - the *Rand Index* and the *Fowlkes-Mallows Index* - have been computed comparing clusters obtained by the k-means and the previously defined clusters within the Data Visualization paragraph. In order to obtain the best configuration, further ten thousand MonteCarlo simulations have been run with the chosen $k = 9$ maximizing the Rand Index and choosing the respective cluster configuration. Table 2 reports the best cluster configuration result with respect to the Rand Index.

Table 2. best external evaluation index.

Rand Index	Fowlkes Index
0.76898	0.644766

4.2. Setting Thresholds

Table 3. thresholds for consumption per square meter and for day/night energy efficiency index.

Building	$^{kWh}/_{year*m^2}$	$EEI_{night/day}$
Scientific depts without lab	$30 - 50$	$0.8 - 1.1$
Scientific depts with lab	$70 - 110$	$1.1 - 1.9$
Humanities depts	< 50	$0.6 - 1.1$
Agrarian depts	$20 - 70$	$1.5 - 2.5$
Medical depts	$50 - 70$	$1.2 - 1.5$
Administrative offices	< 50	$0.4 - 1$

Starting from the Parallel Coordinates graph we defined alert thresholds for the main six clusters - i.e. scientific depts. (without lab.), scientific depts. (with lab), humanities, agrarian and medical depts. and administrative offices. Results and alert thresholds are reported in Tab. 3 with respect two main attributes $EEI_{year,kWh,night/day}$ and $^{kWh}/_{year*m^2}$. We don't report absolute energy consumption per year because it is not interesting as a general index for energy efficiency. Tab. 3 shows that clusters corresponding to scientific depts.. (with lab.), agrarian and medical depts. have an high day/night energy efficiency index, as expected. Scientific depts. (with lab.) shows a higher energy consumption per meter square with respect to agrarian and medical depts. and in general with respect to all other clusters. Administrative offices, scientific depts. (without lab.) and humanities depts.., instead, have a common behaviour with low $^{kWh}/_{year*m^2}$ and $EEI_{year,kWh,night/day}$. Scientific depts. (without lab), generally, present a slightly higher energy consumption at night.

4.3. Monitoring Trends

The final step of the presented process is based on an application of the parallel coordinates method. In this case, we plot different annual energy consumptions on a different axis (each axis represents a different year) where only one attribute may be plotted. This tool, shown in Fig. 4 allows to visualize the historical trend of a chosen energy efficiency index. A useful feature, is the possibility to highlight simultaneously various buildings, in order to observe their historical trends. By simply hoovering the mouse on each polyline, the building energy consumption for the chosen year is shown. By clicking on it, the polyline is highlighted as seen in Fig. 4, where "Biblioteca Dip. Scienze Filologiche" (yellow) and the "Rettorato" (red) stand out. The tool here described is publicly available at `https://goo.gl/YuPTRB`.

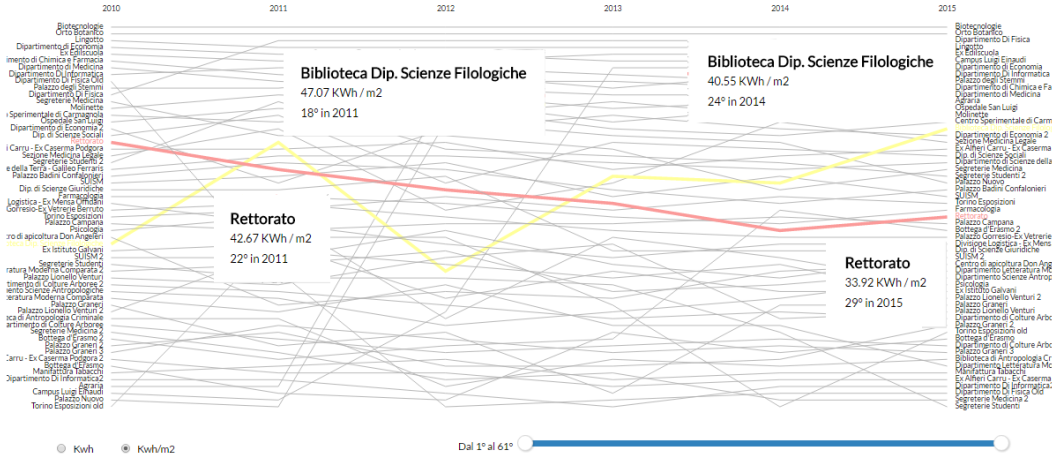


Figure 4. Interactive Data Visualization tool to monitor historical trends based on the Parallel Coordinates method.

5. Discussion

The first aim of this paper was to determine a process to set general hypothesis on building clusters with respect to energy efficiency indices. A clusters hypothesis has been previously stated relying on the background knowledge of the energy management staff at the University of Turin. We envisage this step as a limit of this study, since it requires a preliminary effort by a human task force that can be not always reliable, available, competent or even present. However, the time required in this phase is widely compensated by the easiness of the subsequent steps and the replicability of the monitoring phase in each institution able to offer at least the energy bill data source.

The clusters hypothesis has been made based on main buildings function and then it has been verified via to two methodologies for the identification of buildings clusters: a data visualization approach and a clustering algorithm.

The data visualization approach allowed to recognize the validity of the clusterization hypothesis. In fact, after labelling each building with a precise function, it is possible to match each building within a precise cluster, straightforward (via the brush function). In this way, it is possible to immediately identify outliers and set rough alert thresholds as described in sec. 4.2 and in Tab. 3.

This method made us identify some outliers in the Unito case study. For instance, the *Physics Dept.* and the *Biotechnology Dept.* are two outliers within the cluster "Scientific Depts. - with laboratory". High consumption per square meter and high day/night energy efficiency index are due both to large IT centres and electric chillers running 24/24h. Within the "Agrarian depts." cluster, the *botanical garden* is another outlier, with its very high consumption per square meter. The *Agrarian Campus* has been identified as an outlier, too, with respect to its annual energy consumption. Looking into that, one can infer that since it hosts many thousands of students and very specific function related to field

experiments and greenhouses maintenance, its energy behaviour must be different and must be treated differently. Within the "Medical depts." cluster, the *Dental School* (named “Lingotto” within graphs from the building it is hosted by) and the *legal medicine section* (named Sezione Legale Medicina) are outliers compared to an average energy consumption or a day/night energy efficiency index. Again, a more detailed data source analysis reveals that the Dental School lies in a much bigger complex, the *Lingotto* site, provoking an increase in HVAC use, summed to a high number of dental and technical machinery. As for the legal medicine section, the reason of the high night consumption lies on the morgue and the mortuary rooms, asking for a constant air conditioning system, very costly especially during spring and summer seasons. Within the "Humanities Depts." cluster, there are two outliers, the *Social Science Dept.* and the *Psychology Dept.*, with respect to the day/night index: the reasons of this anomalous consumption is still under studying at the Unito’s facility management office after a signalling coming from this work. *Palazzo Nuovo* has one of the highest number of students and classrooms within the same building, thus explaining its higher energy request. Within the "Scientific Depts - without laboratory", two outliers emerge. The *Management Dept* (named “Dipartimento di Economia”) and *Torino Esposizioni*. The first one has a high consumption per square meter and an high annual consumption because the energy meter counts also the consumption of the Regional IT center, while “Torino Esposizioni” has a very high day/night index because of the secondary function of the building (art exhibitions, fairs and other types of events). Finally, the "Administrative offices" cluster has three outliers: *Palazzo degli Stemmi*, *Manifattura Tabacchi* and *Students’ secretariat*. These three buildings have an high day/night index due to different reasons. The first one is the main building for the technical directions of the University and it hosts a lot of IT server of the University. Other reasons are under investigation. The other two buildings, instead, are two multifunctional buildings hosting public events for the City of Turin.

6. Conclusion

To conclude, this data visualization approach offers a simple way to identify outliers, but the reasons of the inefficiency have to be explained with a deeper analysis, scouting via Google Maps or the facility management office further features that did not emerge during the preliminary labelling phase. As a methodological caveat, this approach reveals outliers within clusters defined ex-ante: therefore, every multifunctional cluster is shown as an outlier of its own cluster, and that can be a limit if a cluster is the result of a preliminary wrong human inference. However, Data Viz techniques revealed to be very useful to explore quickly and simply a large buildings’ stock, identifying the worst efficient buildings and separate their distinct functions.

Secondly, a clustering algorithm has been used in order to test the initial hypothesis. The test was made exploiting two external indices - i.e. the *Rand Index* and the *Fowlkes-Mallows Index* - comparing the clusters configuration hypothesis (hp0) and the obtained clusters thanks to the k-means algorithm. The obtained clusters configuration with $k = 9$ may be compared with the clusters hypothesis (Rand Index = 0.76898). K-means, due to its algorithm basic principles, as many other clustering algorithm is strongly affected by local optimum and outliers. In fact, with a deeper analysis on clusters details, k-means algorithm is able to well-identify outliers - e.g. Management Dept., Biotechnology Dept. or Agrarian Campus - but it recognizes some clusters without physical explanation due to local optimum. For instance, the *Department of Arboree Cultures* (hp0: Agrarian Depts. cluster) and the *Manifattura Tabacchi* (hp0: administrative offices cluster) or the *Don Angeleri Beekeeping Center* (Agrarian Depts. cluster) and the *Psychology Dept.* (hp0: Humanities Depts. cluster) always lye within the same cluster without any other point because of they have a very common energy consumption behaviour. The three clusters - i.e. hp0: administrative offices, humanities depts. and scientific depts. (without laboratories) - are mixed together in only two clusters. Scientific Depts. (with laboratories) cluster is well-recognized loosing one of the outliers described in data visualization approach, the Biotechnology Dept., and gaining two outliers from other clusters, the *Botanical Garden* and the *Dental School*. Many outliers, identified in the data viz approach, are aggregated into the same cluster - e.g. *Manifattura Tabacchi*,

Torino Esposizioni, *legal medicine section*, *Social Science Dept* and *Students' secretariat*. This behaviour reveals that a possible new cluster hypothesis should include a *multifunctional building* cluster. Finally, the two main campuses *Campus Luigi Einaudi* and the *Agrarian Campus* are always grouped together, representing a reasonable choice. The *Management Dept.*, outlier within the Scientific depts. (without lab) cluster, and the *Biotechnology Dept.* are clustered alone. In conclusion, k-means clustering algorithm recognizes very accurately the main clusters – identified as campuses, service industry buildings and Scientific depts. - confirming our initial hypothesis but it is not able, as expected, to recognize slight differences between Humanities depts., Scientific Buildings (without lab) and Administrative Offices.

Results revealed also that clustering algorithms - k-means in our case - cannot be exploited to design useful clusters depending on building functions, except for some macro clusters like tertiary service buildings or campuses and scientific buildings. Moreover, they pointed out how the most interesting part of information in energy efficiency analysis is lost. In fact, data analysts or energy managers are usually interested in inefficient buildings, thus in outliers with respect to their cluster, even when clustering algorithms tend to aggregate outliers in the wrong cluster. This makes a humanised process always necessary and not replaceable. At city level, such data driven tool requires a large penetration of metering systems and possibilities to explore private data of the entire building stock; these conditions are still not easily accessible but combined techniques need to be taken into account for future researches to achieve the desired level of granularity in the data source. Of course, identifying and removing causes of abnormal energy use ensures a more efficient environment and not just in terms of the building energy costs, talking about university campus cases. With our tool, the algorithms applied appears computationally efficient and robust, therefore, they can be easily integrated into existing university campus building energy management and warning systems. Of course, further work is needed to build on this clustering technique to provide additional dataset for training the algorithm, as well as language processing tools for automated analysis of metered building / energy bills data.

Acknowledgments: The investigation has been done under the research fellowship funded by the Fondazione Goria in Italy within the program "Talenti per la Società Civile".

Author Contributions: Dario Cottafava was the principal investigator. Giulia Sonetti contributes in findings proper energy efficiency indices. Paolo Gambino supervised the work as scientific supervisor of the Energy Working Group of the Unito Green Office and Andrea Tartaglino, as Energy Manager of the University of Turin, helped to verify obtained results and to identify proper explanations for the outliers.

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

References

1. Powell, J.B. Green Building Services. *Journal of International Commerce and Economics* **2015**.
2. Wilkinson, P.; Smith, K.; Beevers, S.; Tonne, C.; Oreszczyn, T. Energy, energy efficiency, and the built environment. *Lancet* **2007**, *370*, 1175–1187.
3. Newman, P. The environmental impact of cities. *Environment and Urbanization* **2006**, *18*, 275–295. doi:10.1177/0956247806069599.
4. Staff, I.E.A. *Transition to Sustainable Buildings: Strategies and Opportunities To 2050*; Organization for Economic Cooperation and Development: Paris, 2013. doi:http://dx.doi.org/10.1787/9789264202955-en.
5. Lombardi, P.; Trossero, E. Beyond energy efficiency in evaluating sustainable development in planning and the built environment. *International Journal of Sustainable Building Technology and Urban Development* **2013**, *4*, 274–282. doi:10.1080/2093761X.2013.817360.
6. Brandon, P.S.; Lombardi, P.; Shen, G.Q. *Future challenges in evaluating and managing sustainable development in the built environment*; John Wiley & Sons, 2017.
7. Bakar, N.; Hassan, M.Y.; Abdullah, H.; Rahman, H.; D., P.M.; Hussin, F. Sustainable energy management practices and its effect on EEI: a study on university buildings. Global engineering, science and technology conference; , 2013.
8. Moghimi, S.F.A.; Mat, S.; Lim, C.; Salleh, E.; Sopian, K. Building energy index and end-use energy analysis in large-scale hospitals case study in Malaysia. *Energy Efficiency* **2014**, *7*, 243–256.

9. González, A.B.R.; Díaz, J.J.V.; Caamano, A.J.; Wilby, M.R. Towards a universal energy efficiency index for buildings. *Energy and Buildings* **2011**, *43*, 980 – 987. doi:https://doi.org/10.1016/j.enbuild.2010.12.023.
10. Yun, G.; Steemers, K. Behavioural, physical and socio economic factors in household cooling energy consumption. *Applied Energy* **2011**, *88*, 2191–2200.
11. 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. IEEE, 2007.
12. Ferrer-Balas, D.; Lozano, R.; Huisingh, D.; Buckland, H.; Ysern, P.; Zilahy, G. Going beyond the rhetoric: system-wide changes in universities for sustainable societies. *Journal of Cleaner Production* **2010**, *18*, 607 – 610. Going beyond the rhetoric: system-wide changes in universities for sustainable societies, doi:https://doi.org/10.1016/j.jclepro.2009.12.009.
13. Agdas, D.; Srinivasan, R.; Frost, K.; Masters, F. Energy Use Assessment of Educational Buildings: Toward a Campus-wide Sustainable Energy Policy. *Sustainable Cities and Society* **2015**, *17*.
14. Chung, M.; Rhee, E. Potential opportunities for energy conservation in existing buildings on university campus: A field survey in Korea. *Energy and Buildings* **2014**, *78*, 176–182.
15. Escobedo, A.; Briceño, S.; Juárez, H.; Castillo, D.; Imaz, M.; Sheinbaum, C. Energy consumption and GHG emission scenarios of a university campus in Mexico. *Energy for Sustainable Development* **2014**, *18*, 49 – 57. doi:https://doi.org/10.1016/j.esd.2013.10.005.
16. Evans, J.; Jones, R.; Karvonen, A.; Millard, L.; Wendler, J. Living labs and co-production: university campuses as platforms for sustainability science. *Current Opinion in Environmental Sustainability* **2015**, *16*, 1 – 6. Sustainability science, doi:https://doi.org/10.1016/j.cosust.2015.06.005.
17. Robinson, O.; Kemp, S.; Williams, I. Carbon management at universities: A reality check. *Journal of Cleaner Production* **2014**, *106*.
18. del Mar Alonso-Almeida, M.; Marimon, F.; Casani, F.; Rodriguez-Pomeda, J. Diffusion of sustainability reporting in universities: current situation and future perspectives. *Journal of Cleaner Production* **2015**, *106*, 144 – 154. Bridges for a more sustainable future: Joining Environmental Management for Sustainable Universities (EMSU) and the European Roundtable for Sustainable Consumption and Production (ERSCP) conferences, doi:https://doi.org/10.1016/j.jclepro.2014.02.008.
19. Lauder, A.; Sari, R.F.; Suwartha, N.; Tjahjono, G. Critical review of a global campus sustainability ranking: GreenMetric. *Journal of Cleaner Production* **2015**, *108*, 852 – 863. doi:https://doi.org/10.1016/j.jclepro.2015.02.080.
20. of Statistics of China (NBSC), N.B. China Statistical Yearbook. Technical report, China Statistics Press, China, 2012.
21. Shriberg, M. Institutional assessment tools for sustainability in higher education: Strengths, weaknesses, and implications for practice and theory. *International Journal of Sustainability in Higher Education* **2002**, *3*, 254–270. doi:10.1108/14676370210434714.
22. Haas, R. Energy efficiency indicators in the residential sector: What do we know and what has to be ensured? *Energy Policy* **1997**, *25*, 789 – 802. Cross-country comparisons of indicators of energy use, energy efficiency and CO2 emissions, doi:https://doi.org/10.1016/S0301-4215(97)00069-4.
23. Jollands, N.; Patterson, M. Four theoretical issues and a funeral: improving the policy-guiding value of eco-efficiency indicators. *International Journal of Environment and Sustainable Development* **2004**, *3*, 235–261. doi:10.1504/IJESD.2004.005074.
24. Sonetti, G.; Lombardi, P.; Chelleri, L. True Green and Sustainable University Campuses? Toward a Clusters Approach. *Sustainability* **2016**, *8*. doi:10.3390/su8010083.
25. Yik, F.; Burnett, J.; Prescott, I. Predicting air-conditioning energy consumption of a group of buildings using different heat rejection methods. *Energy & Buildings* **2001**, *33*, 151–166.
26. Howard, B.; Parshall, L.; Thompson, J.; Hammer, S.; Dickinson, J.; Modi, V. Spatial distribution of urban building energy consumption by end use. *Energy and Buildings* **2012**, *45*, 141 – 151. doi:https://doi.org/10.1016/j.enbuild.2011.10.061.
27. Yang, C.; Létourneau, S.; Guo, H. Developing Data-driven Models to Predict BEMS Energy Consumption for Demand Response Systems. Modern Advances in Applied Intelligence; Ali, M.; Pan, J.S.; Chen, S.M.; Horng, M.F., Eds.; Springer International Publishing: Cham, 2014; pp. 188–197.

28. Hong, T.; Yang, L.; Hill, D.; Feng, W. Data and analytics to inform energy retrofit of high performance buildings. *Applied Energy* **2014**, *126*, 90 – 106. doi:https://doi.org/10.1016/j.apenergy.2014.03.052.
29. Yalcintas, M. An energy benchmarking model based on artificial neural network method with a case example for tropical climates. *International Journal of Energy Research* **2006**, *30*, 1158–1174. doi:10.1002/er.1212.
30. M., Y.; Aytun, O. An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database. *International Journal of Energy Research* **2007**, *31*, 412–421.
31. Fan, C.; Xiao, F.; Wang, S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy* **2014**, *127*, 1 – 10. doi:https://doi.org/10.1016/j.apenergy.2014.04.016.
32. Inselberg, A. Multidimensional Detective. Information Visualization. IEEEE, 1997.
33. Card, S.K.; Mackinlay, J.; Shneiderman, B. Readings in Information Visualization: Using Vision to Think. *Morgan Kaufman* **1999**.
34. NIST-SEMATECH. E-Handbook of Statistical Methods, 1997.
35. Bostock, M.; Ogievetsky, V.; Heer, J. D3: Data-Driven Documents. *IEEE Trans Vis Comput Graph* **2011**, *12*, 2301–9.
36. Bezanson, J.; Edelman, A.; Karpinski, S.; Shah, V.B. Julia: A fresh approach to numerical computing. *arXiv:1411.1607* **2014**.
37. Keim, D. Visual techniques for exploring databases. Technical report, NIST, 2003.
38. Inselberg, A. The plane with parallel coordinates. *Visual Computer* **1985**, pp. 1:69–97.
39. Feiner, S.; Beshers, C. Worlds within worlds: metaphors for exploring n-dimensional virtual worlds. 3rd annual ACM SIGGRAPH symposium on User interface software and technology, 1990, pp. 76–83.
40. Cleveland, W. *Visualizing Data*; Hobart Press, 1993.
41. Borg, I.; Groenen, P.J.F. Modern Multidimensional scaling: theory and Applications. *Visual Computer* **2005**.
42. Keller, P.R.; Keller, M.M. Visual Cues-Practical Data Visualization. *IBM Systems Journal* **1993**, *33*.
43. 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. 48th International Conference AICARR, 2011, pp. 547–559.
44. Cottafava, D.; Gambino, P.; Baricco, M.; Tartaglino, A. Multidimensional analysis tools for energy efficiency in large building stocks. 12th Conference on Sustainable Development of Energy, Water and Environment Systems, 2017.
45. Jain, A.; Dubes, R. *Algorithms for clustering data*; Prentice-Hall: Upper Saddle River, 1988.
46. Xu, R.; Wunsch, D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks* **2005**, *16*, 645–678. doi:10.1109/TNN.2005.845141.
47. Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science* **2015**, *2*, 165–193. doi:10.1007/s40745-015-0040-1.
48. Kassambara, A. *Practical Guide To Cluster Analysis in R*; STHDA, 2017.
49. Maulik, U.; Bandyopadhyay, S. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2002**, *24*, 1650–1654. doi:10.1109/TPAMI.2002.1114856.
50. Starczewski, A.; Krzyżak, A. Performance Evaluation of the Silhouette Index. Artificial Intelligence and Soft Computing; Rutkowski, L.; Korytkowski, M.; Scherer, R.; Tadeusiewicz, R.; Zadeh, L.A.; Zurada, J.M., Eds.; Springer International Publishing: Cham, 2015; pp. 49–58.
51. Rand, W.M. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* **1971**, *66*, 846–850, [https://www.tandfonline.com/doi/pdf/10.1080/01621459.1971.10482356]. doi:10.1080/01621459.1971.10482356.
52. Kosub, S. A note on the triangle inequality for the Jaccard distance. *CoRR* **2016**, *abs/1612.02696*, [1612.02696].
53. Fowlkes, E.B.; Mallows, C.L. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association* **1983**, *78*, 553–569, [https://www.tandfonline.com/doi/pdf/10.1080/01621459.1983.10478008]. doi:10.1080/01621459.1983.10478008.

54. Nagpal, A.; Jatain, A.; Gaur, D. Review based on data clustering algorithms. 2013 IEEE Conference on Information Communication Technologies, 2013, pp. 298–303. doi:10.1109/CICT.2013.6558109.
55. Ahmad, A.; Dey, L. A K-mean Clustering Algorithm for Mixed Numeric and Categorical Data. *Data Knowl. Eng.* **2007**, *63*, 503–527. doi:10.1016/j.datak.2007.03.016.
56. Macqueen, J. Some methods for classification and analysis of multivariate observations. In 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.
57. Park, H.; Jun, C. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications* **2009**, *36*, 3336–3341. doi:doi:10.1016/j.eswa.2008.01.039.
58. Kaufman, L.; Rousseeuw, P. *Partitioning around medoids (program pam)*; Wiley, Hoboken, 1990; pp. 126–160.
59. Ng, R.T.; Jiawei, H. CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering* **2002**, *14*, 1003–1016. doi:10.1109/TKDE.2002.1033770.
60. Kaufman, L.; Rousseeuw, P. *Partitioning around medoids (program pam)*; Wiley, Hoboken, 1990; pp. 68–120.
61. Guha, S.; Rastogi, R.; Shim, K. CURE: An Efficient Clustering Algorithm for Large Data sets. Published in the Proceedings of the ACM SIGMOD Conference, 1998.
62. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: an efficient data clustering method for very large databases. In Proc. of the ACM SIGMOD Intl. Conference on Management of Data (SIGMOD, 1996, pp. 103–114.
63. Karypis, G.; Han, E.H.; Kumar, V. Chameleon: hierarchical clustering using dynamic modeling. *Computer* **1999**, *32*, 68–75. doi:10.1109/2.781637.
64. Ketchen, J.D.; Shook, C.L. The application of cluster analysis in strategic management reasearch: an analysis and critique. *Strategic Management Journal* **1996**, *17*, 441–458. doi:https://doi.org/10.1002/.
65. Thorndike, R.L. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276. doi:10.1007/BF02289263.
66. Pollard, K.S.; Van Der Laan, M.J. A method to identify significant clusters in gene expression data. SCI (World Multiconference on Systemics, Cybernetics and Informatics), 2002, Vol. 2, pp. 318–325.
67. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2001**, *63*, 411–423.
68. Sheikholeslami, G.; Chatterjee, S.; Zhang, A. Wavecluster: A multi-resolution clustering approach for very large spatial databases. VLDB, 1998, Vol. 98, pp. 428–439.
69. Smyth, P. Clustering Using Monte Carlo Cross-Validation. Kdd, 1996, Vol. 1, pp. 26–133.
70. Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **2003**, *52*, 91–118.
71. Roth, V.; Lange, T.; Braun, M.; Buhmann, J. A resampling approach to cluster validation. Compstat. Springer, 2002, pp. 123–128.
72. Wang, J. Consistent selection of the number of clusters via crossvalidation. *Biometrika* **2010**, *97*, 893–904. doi:10.1093/biomet/asq061.
73. Cottafava, D.; Gambino, P.; Baricco, M.; Tartaglino, A. Energy efficiency in a large university: the UniTo experience. Sustainable Built Environment. Towards Post Carbon Cities, 2016, pp. 92–101.

Sample Availability: Datasets and further information are available from the authors.