A data-mining approach for energy behavioural analysis to ease predictive modelling for the smart city

L.C. Tagliabue [a], S. Rinaldi [b], M. Favalli Ragusini [c], G. Tardioli [c,d], A.L.C. Ciribini [a]

[a] Department of Civil, Environmental, Architectural Engineering and Mathematics, University of Brescia, Italy
[b] Department of Information Engineering, University of Brescia, Italy
[c] Integrated Environmental Solutions (IES) R&D, Glasgow, UK
[d] School of Mechanical & Materials Engineering, University College Dublin, Dublin, Ireland

E-mail: lavinia.tagliabue@unibs.it, stefano.rinaldi@unibs.it, mario.ragusini@iesve.com,
giovanni.tardioli@iesve.com, angelo.ciribini@unibs.it

Abstract –

Analyzing urban districts to promote energy efficiency and smart cities control could be very complex as big data have to be analyzed filtered to discover unpredictable patterns. Using clustering method, specifically K-means algorithm, allows to create an energy profiling characterization of urban district models with multiple advantages: as first, large quantity of data can be managed and synthesized, easing the creation of algorithm patterns that could be replicable. Thus, it is possible to operate in a large scale and in a small scale in the same time, choosing the level of detail that is more appropriate for the specific analysis. In the large scale, the disadvantages are the dependence from data, i.e. if there are missing input values, it is hard to rebuild them because of the quantity of data. Missing values can confuse the analysis because scripts cannot identify the entire row missing it. Working with clustering analysis it is thus useful when large amount of data should be organized and interpreted and the technique can help the planner to make faster the analyses process. The research aims at demonstrate the efficiency of clustering methods when adopted for energy consumption issues at city level. In the paper, the clustering process concerning building energy profiles of a European city for the identification of building models is described. This means that an energy template on urban scale is used and clusters are applied on energy profiles based on architectural and energy similarity in order to find representative models. In particular, the study is focused on the relationships between building characteristics and actual building energy profiles.

Keywords –

Clusterization, behavioural modelling, smart city data mining, energy profiling

1 Introduction

In the last years, energy analysis has been progressively boosted to reach high efficiency in buildings, increasing savings and reducing CO_2 emissions according with EU objectives. EU government has set three targets to be achieved by the 2020: 20% reduction of greenhouse gas emissions (starting from 1990 value), 20% increase of EU energy from renewables, 20% improvement in energy efficiency [1].

The building sector is responsible of 40% of European Union’s total energy consumption.

The population growth and the strong urbanization process showed in 2014 that the 54% of the world’s population is living in urban areas. This proportion is expected to rise to 66% by 2050. Moreover, an increase of 90% is expected to be concentrated in Asia and Africa, according UN report [2]; this situation requires a more consistent energy management in order to reduce wastages and to increase comfort. This goal could be reached by the use of energy monitoring devices that can help to optimize energy consumption. Several studies show that is essential to analyze building energy profiles in order to understand how the occupancy variation affects the energy behavior [3][4] and data gathered are the key factor to understand the consumption distribution and thus to manage the smart city [5][6].

Nowadays, data mining is a powerful framework to promote consciousness and thus achieve strategies efficiency. Cooperation between different clusters of consumers, prosumers and energy production spots and plants [7], methods to analyze big and giant data are crucial to develop models [8] to represent, predict and renovate the cities through IoTed solutions and digital based technologies [9][10]. Digitalization is changing the way we approach the concerns in many fields however the smart city and the energy management is one of the first and exploited sectors in which digital models and predictive as well simulators are adopted extensively since many years. The digital revolution requires a high level of specialization in many fields such as data analysis and coding, computing, data engineering, machine learning, artificial intelligence (AI).

Digitalization is basically dependent on data and for that reason, IoT is increasing and data are gathered in every system up to redundancy. Nonetheless, the amount of data is not a guarantee of amount of information because of unstructured data or data organized in different ways compared to readable ones is leading to a big amount of unused data (90%). The initial vision of the final use and meaning and the correct plan of data collection should drive the "thirst" of data to make them useful. In AEC sector, the adoption of computers is increasingly fundamental as more and more data have to be analyzed and knowledge extracted to outline the predictive models and promote a robust analytical assessment.

2 Methodology

The methodology adopted is based on mathematical methods that are used to operate a data mining process and clusterization of big data. In the following subsections, the specific methods considered together with advantages and weaknesses are described before to go in deep with the methodology and application to the case study of the present research.

2.1 Clustering methods

With the grown quantity of data, clustering methods have been developed to make easier the data mining procedure using many clustering methods such as Hierarchical, K-means, K-medoid, etc. Clustering is an unsupervised learning task that aims at decomposing a given set of objects into subgroups or clusters based on similarity. The goal is to divide the data set in such a way that objects belonging to the same cluster are as similar as possible, whereas objects belonging to different clusters are as dissimilar as possible. Cluster analysis is primarily a tool for discovering previously hidden structure in a set of unordered objects [11].

Clusters are useful for creating and analyzing building energy profiles in order to organize data collected into groups easing to analyze and evaluate them also increasing the relevance of measures and indicators. For example, data could concern the energy consumption related to the use of kitchen and laundry appliances, electronic devices such as TVs or computers. Clusters can be prepared based on the different kind of appliances or based on the different users' behavior in the analyzed buildings [12] and collected through smart meters [13]. Based on specific induction principle many algorithms have been developed for clusterization and the main approaches are hierarchical methods and partitioning methods.

2.1.1 Hierarchical methods

These methods could be called top-down or bottom-up methods since they identify clusters following these two main directions. There are two main types of hierarchical methods and the difference sets in the starting point of the analysis:

- *Agglomerative hierarchical clustering* starts with single objects and each object represents a cluster. Then clusters are merged until a default structure is obtained.
- *Divisive hierarchical clustering* begins with one single cluster that contains all the objects and the subdivision in more clusters occurs successively.

The result of the analysis is a dendrogram representing all different considered groups, which are subdivided into levels (Figure 1).

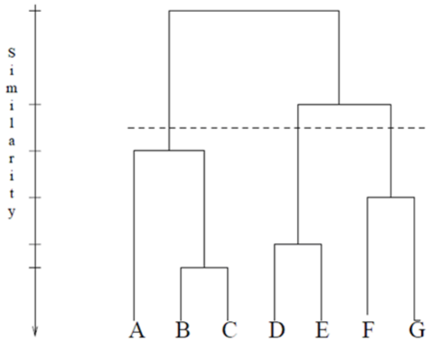


Figure 1. Dendogram: example of hierarchical results representation.

Checking the desired level of similarity, the clustering of objects could be obtained.

In the hierarchical method, more subdivisions that are possible come from the way of the similarity measure of clusters:

- *Single-link clustering* considers the distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster. If it is based on similarities, the similarity between couples of clusters is considered equal to the greatest similarity from any member of one cluster to any member of the other cluster.
- *Complete-link clustering* considers the distance between two clusters to be equal to the longest distance from any member of one cluster to any member of the other cluster.

- *Average-link clustering* considers the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster.

The hierarchical method has the disadvantage of the inability to scale because of the time complexity that is non-linear with the number of objects. The method also has a high degree of rigidity.

2.1.2 Partitioning methods

Partitioning methods consist in move instances from one cluster to another, starting from an initial partitioning. The characteristic of this method is that the number of clusters has to be pre-set. To grasp the optimization an exhaustive enumeration process of all possible partitions is required [14]. There are different types of partitioned-based clustering. The most frequently used are the error minimization algorithms, which works perfectly with compact and isolated clusters.

When the distance of each instance to its representative value is measured, a certain error occurs. These algorithms try to minimize that error. The Sum of Squared Error (SSE) is the most known method and it measures the squared Euclidean distance of instances to their representative values. The most used error minimization algorithm is the K-means algorithm, which divides objects into K clusters, chosen randomly and each cluster is represented by a center or mean. The center is calculated as the mean of all the instances belonging to that cluster:

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_q$$

Where N_k is the number of instances belonging to cluster k and µk is the mean of the cluster k.

The linearity of K-means method is the key of its success over other clustering methods and even if the number of objects is huge, this algorithm is computationally attractive (Figure 2).

The weakness of this method is the choice of the initial number of clusters: the process of clustering is based on this number and the initial choice can change the entire result since there can be differences between global and local minimum.

Another partitioning algorithm that is part of the error minimization algorithms is the K-medoid or PAM (Partition Around Medoids); this algorithm is very similar to the K-means but it differs from the latter since the representation of the different clusters is dissimilar.

The principle is the same: each cluster has a center that, in this case, is the medoid; a medoid is the most centric object in the cluster and it is not influenced by extreme values and for this reason is more solid than the K-means method. Throughout evaluation methods, clustering criteria compute the optimal number of groups for a dataset calculating the index for each case. They measure the compactness of clusters and the homogeneity of them and choose the cluster for each element.

Graph-Theoretic clustering is a method that provides clusters via graphs. Instances are represented as nodes and this method connect all nodes. There is a connection between hierarchical method and the graph theoretic clustering since there is similarity between the hierarchical representation and the Minimal Spanning Tree (MST).

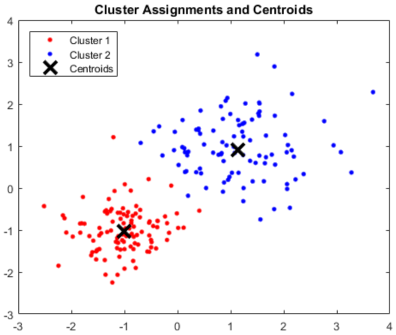


Figure 2. Example of K-means results representation.

2.2 Application in the research

The research consists into analyzing data about energy behavior of a big quantity of buildings gathered by means of smart meters used to register energy consumption during the all year with a step of 15 minutes. Thus, every quarter of hour, smart meters sent the registered data to the central system for monitoring. The analysis is used to identify the specific consumption patterns related to uses and building models. The smart meters have an identification number and for this reason, every smart meter represents the ID (Identification Code/Document) of a specific flat or activity set in a certain building. The research methodology is based on collecting data, sorting data and analyzing them following the scheme (Figure 3).

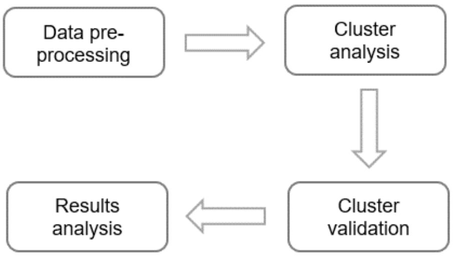


Figure 3. Methodological workflow.

2.3 Data analysis

The methodology start from a first analysis on data metered, after that, the different trends about electrical consumptions have been represent. The representative plots are performed using the ggplot2 R package.

In order to have a clear knowledge about the general trend of data, the smart meters' profiles of the entire year were plotted and compared in order to visualize. the data to be organized to allow the clustering analysis.

With the intention of creating a resume table with mean, median, maximum, minimum and the standard deviation, the main database has been partitioned selecting the unique ID of the smart meters. A table composed by 659 observations (i.e. 659 unique Smart meter IDs in the original database) and n. 6 variables has been created. After the creation of clusters, analysts are able to start the process of clustering validation that is the measure of distances of each data from the center of cluster (centroid) to understand the dissimilarity within each group [15]. This process helps to get a clearer vision about the objects giving a summarizing plot and data tables, which allow comparing results in an efficient way. Than the filtering data procedure started to divide into smaller packages to be handle. Data analyses and cluster methods permits to collect the different information in an easier way than in the past and these methods are suitable for smart cities analysis and strategies [16].

2.4 Cluster analysis

Data analysis allowed collecting various information about thousands of buildings including the geometry, the energy consumptions and the location; hence, it was possible to have insight on the composition of the analysed built environment. The information were about the typology of the analysed buildings such as the use, the year of construction and if changes occurred during the years. Throughout the data, it was possible to perform the analysis on the energy consumption that is strongly influenced by the use of the buildings.

After the data analysis, clusters were organized and the method adopted was the partitioning method with the help of the K-means algorithm.

This process is composed by two steps:

1. A first evaluation by using NbClust R package has been performed. This package provides n. 30 indices for determining the number of clusters and proposes to users the best clustering scheme.
2. The second step was to put the number of clusters that has been identified by the evaluation into the K-means function of R in order to create the subdivision following the similarities of the smart meters' profiles.

As described in section 2, clustering is the partitioning of a set of objects into groups in order to obtain groups with similar objects grouped in the same cluster. Most of the clustering algorithms depend on assumptions with the intention to define the subgroups present in a dataset.

The evaluation process had to tackle difficult problems such as

- the quality of clusters;
- the degree with which a clustering scheme fits a specific data set;
- the optimal number of clusters in a partitioning.

The evaluation was performed with NbClust. All these clustering validity indices combine information about intra-cluster compactness and inter-cluster isolation, as well as other factors, such as geometrical or statistical properties. There are n. 30 index but in the present research, two main indexes has been considered:

- D-index;
- Hubert-index.

2.4.1 D-index

The D-index is based on clustering gain on intra-cluster inertia that measures the degree of homogeneity between the data associated with a cluster. It calculates their distances compared to the reference point representing the profile of cluster that is the cluster centroid in general. The equation [17] representing this index is:

$$w(P^q) = \frac{1}{q}\sum_{k=1}^{q}\frac{1}{n_k}\sum_{x_i \in C_k} d(x_i, C_k)$$

The D index is a graphical method of determining the number of clusters. In the plot of D-index, we seek a significant knee (the significant peak in D-index second differences plot) that corresponds to a significant increase of the value of the measure.

2.4.2 Hubert-index

The Hubert-index *r* statistics is the point serial correlation coefficient between any two matrices. When the two matrices are symmetric, *r* can be written in its raw form as the equation:

$$\mathrm{r(P, Q)} = \frac{1}{N_t}\sum_{\substack{i=1 \\ i<j}}^{n-1} P_{ij}Q_{ij}$$

Where:

- P is the proximity matrix of the data set;
- Q is an n x n matrix whose (i, j) element is equal to the distance between the representative points (vc_i,

vc$_j$) of the clusters where the objects x_i and x_j belong.

High values of normalized *r* statistics indicate the existence of compact clusters. The Hubert index is a graphical method of determining the number of clusters. In the plot of Hubert index, we seek a significant knee that corresponds to a significant increase of the value of the measure i.e. the significant peak in Hubert-index second differences plot [18].

After the evaluation, for the application in the case study (section 4) the NbClust function suggested a number of clusters k=5 plotting two graphs: the first one indicates the D-index values and it is possible to understand the number of cluster from the graph seeing the knee of the trend (Figure 4). The second graphs indicates the second differences D-index values and it suggests that the n. 5 is the ideal number of clusters throughout a peak [19]. With the intention of operating the clustering algorithm, the K-means function has been used setting n. 5 different clusters and using the Hartigan-Wong algorithm [20]. This kind of clustering algorithm determines a cluster assignment of a point peaking it repeatedly. It is based on the sum of squares of errors (SSE) and it searches the optimal within-cluster SSE for assigning an object to the proper cluster.

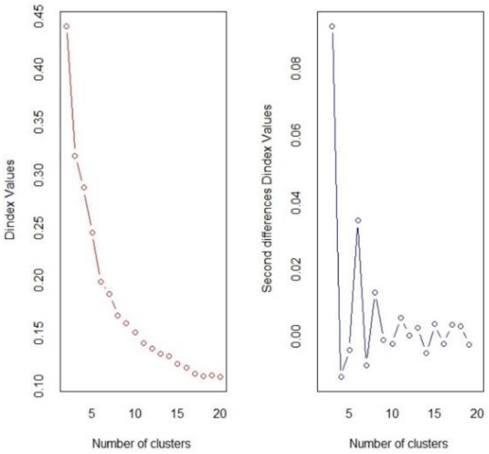


Figure 4. Indications by Hubert and D-index indices for the case studies.

3 Case study

The case study has been a district with 10.000 buildings of a European city where a smart metering strategy has been applied and data collected.

The organization of the buildings' data was based on the ID used in the data processing. After having sorted data, the next stage has been to analyze them trying to find out a significant clustering algorithm. To achieve that, an analysis on the buildings' use has been performed in order to collect Smart Meters with energy values within a correctly sized wide range. Then, a classification of the most frequent destinations of buildings was operated. The most frequent uses in the case study are listed below:

- Banks;
- Coffee shops and restaurants;
- Retirement Homes;
- Municipal Houses
- Residences,
- Private properties and Residential buildings;
- Administration buildings
- Industries and factories.

It is possible to define the type of buildings and the most frequent typology analyzing the residential use: the most recurring was the multifamily houses, followed by the single-family houses. Therefore, the analysis of the multifamily houses was developed since this category had the largest number of samples (Figure 5). Data processing delivers a list collecting the energy consumption value during the entire year identifying day by day, hour by hour, the effective energy usage.

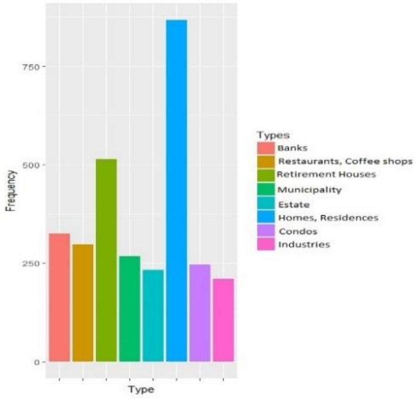


Figure 5. Buildings' use frequency diagram.

The research allowed getting a database composed by 22'994'372 observations with 659 unique Smart Meters IDs. This means that for each ID there are *n* rows indicating day, month, year, hour, and minute for the entire year. The variables of this data frame are n. 9.

This database is significant in order to design energy profiles for each smart meter, to get many other databases analyzing as an example data into workdays or festive days, analyzing month by month the energy use, visualizing systematic trends of these profiles.

4 Results

In order to understanding the results of the clustering analysis, a series of plots have been performed. The first representation of this study consists of plotting singles clusters in order to understand the differences between the groups. In Figure 6 the data for each cluster for one working day are plotted with the representative elements.

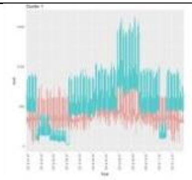
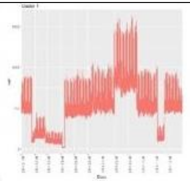
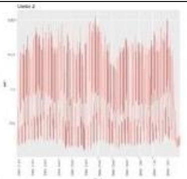
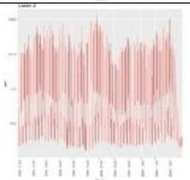
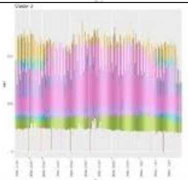
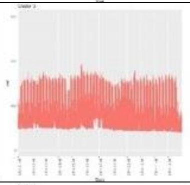
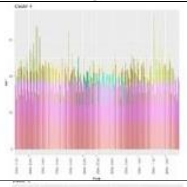
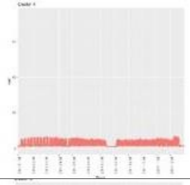
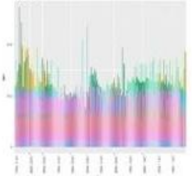
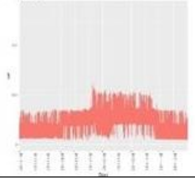
	Clusters	Representative Elements
1		
2		
3		
4		
5		

Figure 6 Working January One day profile.

The energy consumed during a year (2014) is thus plotted for each cluster considering the data time step of 15 minutes. In the following, the clusters are listed and data discusses:

- Cluster 1 is characterized by kWh values ranging between 0 and 1600 kWh. This cluster contains two cases and it has the highest range of kWh values.
- Cluster 2 is the smallest one and it gathers one observation that ranges between about 400 kWh and 1260 kWh with a value of around 860 kWh per year.
- Cluster 3 is composed by 8 objects and it has a range that goes from 0 kWh to almost 300 kWh.
- Cluster 4 contains 618 profiles and it is the largest one. The energy values ranges between 0 and 75 kWh and it is the cluster with the lowest range of values.
- Cluster 5 gathers 30 smart meter's profiles and it has an average value of almost 300 kWh with data ranging between 0 and 280 kWh.

In Table 1 the characteristics values of the clusters are reported. It easy to understand that cluster 4 is the cluster in which residential buildings are gathered and which is the most populated cluster. Cluster 1 could be the industrial buildings one and the municipal house could be cluster 2. The idea is that analyzing the data is not needed to go in deep into the identity of the building (with eventually privacy problems) but is possible to define the values trend and to suggest possible implementation strategies through the analysis of the archetypes, also calculating measure to redefine peaks and trends.

Table 1 Clusters Characteristic values

Characteristics	Clusters				
	1	2	3	4	5
Size	2	1	8	618	30
Mean	12.03	16.47	2.63	-0.13	0.73
Median	11.27	17.93	2.42	-0.13	0.60
Max	13.22	13.52	2.02	-0.14	0.95
Min	0.91	23.38	2.61	-0.08	0.22
SD	12.75	13.85	1.68	-0.13	0.89

In Figure 7, the whole profiles are shown to the left side, a working day (Monday) data are plotted to the right side.

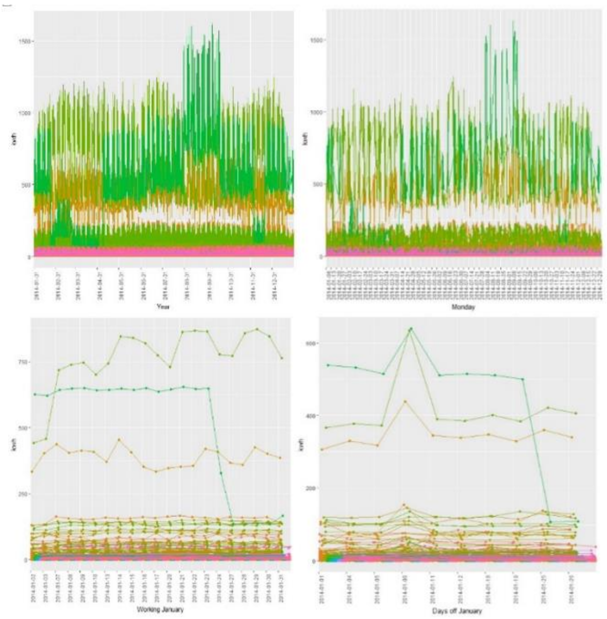


Figure 7 (above) One year smart meters profile and Monday profile and (below) week working day (January) and weekend daily profile.

It is possible to check that the different profiles are defining the city uses and buildings in which the energy consumption is concentrated.

There are the big consumers (Cluster 1 and 2) and all the other profiles with less consumption each but with a great number of units and variability due to occupancy distribution and different behaviors.

Cluster 4 is thus the largest with 618 profiles and this represents the housing stock in the sample used as a case study. For the main cluster 4 is thus shown in Figure 8 for the working day in January and the week end day in the same month.

In the left side diagrams, the yearly-recorded data are plotted and to the right side the daily profile are shown.

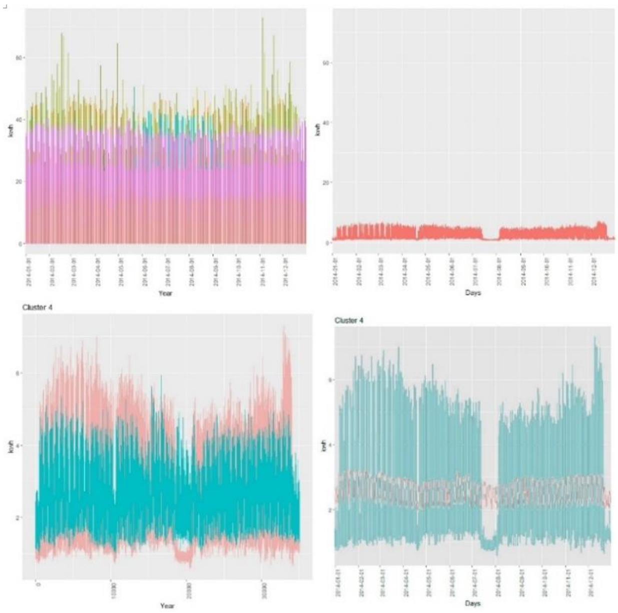


Figure 8 Week working day profile (January) and weekend daily profile.

The representative element of the Cluster represents the profile of a building and an ideal element that has been created through the average and the weighted average of the cluster.

The research showed that the weighted average is most suitable to define this representative profile because as multiple elements are used in the weighted average, the ideal centroid moves away from the representative element.

5 Conclusion

Hence, the digitalization has brought a strong advantage that is the possibility of managing a larger database that can make the pre-analysis process and that can analyse itself faster with a very significant improvement for the time management. A less time expensive but detailed analysis allows to the reduce costs which is one of the main aspects considered by the clients. Europe and the entire world is getting effort in order to achieve important results regarding the reduction of GHG emissions to decrease the global warming effect. The environmental benefits that energy savings through technology advancement are bringing and the way digital revolution is helping the management of energy sources are very important to achieve these goals.

There is a transition from old business models to the new ones that means that private sector and utilities companies are investing on renewable energy and new technologies, for example through smart meters. Smart meters (SM) are very helpful in order to collect a very large quantity of data and since the IoT (Internet of Things) is emerged, massive amounts of data are available and they are easier to manage. SM are bringing many benefits for consumers and for utilities: customers can be constantly informed about their consumptions and the way of communication can be remotely controlled with historical data and locally the consumers are informed by real-time data with a significant improvement of customers' knowledge about their energy costs and carbon emission. These devices allow the electrical appliances to be automatically controlled and this feature is very interesting since consumers can easily manage the electrical system selecting days off where the consumptions can be reduced and, oppositely, the days when users request more power to the system.

In this way, waste of energy can be reduced. For utilities, the advantages are that they can influence the energy profiles of their users; they can have a reduction in 'costs to serve' rationalizing their services and they are able to sensitize customers about global warming issue.

Smart Meters and IoT bring a large quantity of data, as we said, and clustering methods have been and will be critical in order to manage in a better way this data [21]. Clustering method has many advantages: first, it increase productivity through specialized inputs, access to information, synergies [22]. The data management through clustering methods allows creating an energy profiling characterization of urban building models suitable to study and define balance strategies and flatting peaks procedures coupled with policies on energy sources to be adopt considering IoT and Smart Management of the city. The advantages of using clustering methods, specifically K-means algorithm, are that it eases to manage a large quantity of data, synthesized them and create an algorithm pattern that could work for further different cases and thus could be transposed. Therefore, it is possible to operate in a large scale and in a small scale in the same time, choosing the level of details which is more appropriate for the analysis and the calculation useful for energy planning and to support and explain future development.

References

[1] Chen D., Zhang S., Xue Q., Brief investigation of sensor technology and data analysis in building energy management. *28th Chinese Control and Decision Conference (CCDC)*, 2016.

[2] United Nations, *World Urbanization prospects*, Published by the United Nations, 2014, ISBN 978-92-1-151517-6.

[3] Hong T., D'Oca S., Turner W.J.N., Taylor-Lange S.C., An ontology to represent energy-related occupant behavior in buildings. Part I: Introduction to the DNAs framework, *Building and Environment*, Volume 92, October 2015, Pages 764-777.

[4] Hong T., Taylor-Lange S.C., D'Oca S., DaYan, Corgnati S.P., Advances in research and applications of energy-related occupant behavior in buildings, *Energy and Buildings*, Volume 116, 15 March 2016, Pages 694-702.

[5] Calvillo C.F., Sánchez-Miralles A., Villar J., Energy management and planning in smart cities, *Renewable and Sustainable Energy Reviews*, Volume 55, March 2016, Pages 273-287.

[6] Bianchini, D., De Antonellis, V., Melchiori, M., Bellagente, P., Rinaldi, S., Data management challenges for smart living, 2018 Lecture Notes of the Institute for Computer Sciences, *Social-Informatics and Telecommunications Engineering*, LNICST, 189, pp. 131-137.

[7] Pan G., Qi G., Zhang W., Li S., Wu Z., Yang L.T., Trace analysis and mining for smart cities: issues, methods, and applications, *IEEE Communications Magazine*, Vol.: 51, n. 6, June 2013, pp.: 120- 126.

[8] Zhou K., Fu C., Yang S., Big data driven smart energy management: From big data to big insights, *Renewable and Sustainable Energy Reviews*, Volume 56, April 2016, pp. 215-225.

[9] Chen F., Deng P., Wan J., Zhang D., Vasilakos A.V., and Rong X., Data Mining for the Internet of Things: Literature Review and Challenges, *International Journal of Distributed Sensor Networks*, Volume 2015, Article ID 431047, 14 pages.

[10] Pasini, D., Mastrolembo Ventura, S., Rinaldi, S., Bellagente, P., Flammini, A., Ciribini, A.L.C., "Exploiting internet of things and building information modeling framework for management of cognitive buildings", *in Proc. of IEEE International Smart Cities Conference*, ISC2 2016 Trento, Italy.

[11] Goia A., May C., Fusai G., Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting* 26, (2010) 700-711.

[12] Alahakoon D., Yu X., Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey, *IEEE Transactions on Industrial Informatics*, Vol.: 12, n: 1, Feb. 2016, pp. 425- 436.

[13] Dede' A., Della Giustina D., Rinaldi S., Ferrari P., Flammini A., Smart meters as part of a sensor network for monitoring the low voltage grid, *2015 IEEE Sensors Applications Symposium (SAS)*, Zadar, Croatia, April 13-15, 2015, pp. 276-281, ISBN 978-1-4799-6117-7, DOI 10.1109/SAS.2015.7133616.

[14] Domeniconi C., Al-Razgan M., Weighted cluster ensembles: Methods and analysis, Transactions *on Knowledge Discovery from Data (TKDD)*, Volume 2 Issue 4, January 2009, Article No. 17.

[15] Benitez I., Quijano A., Diez J.L., Delgado I., Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. *Eletrical Power and Energy Systems*, 55, (2014) pp. 437-448.

[16] Charrad M., Ghazzali N., Boiteau V., Niknafs A., NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set, *Journal of Statistical Sofwtare*, Volume 61, Issue 6, 2014.

[17] Morissette L., Chartier S., The k-means clustering technique: General considerations and implementation in Mathematica, *Tutorials in Quantitive Methods for Psychology*, Volume 9(1), p. 15-24, 2013

[18] Ma J., Cheng J.C.P. Estimation of the building energy use intensity in the urban scale by integrating GIS and big data thechnology, *Applied Energy*, 2016, 183, pp 182-192.

[19] Kabacoff R.I., R in Action, *Data analysis and graphics with R*, Shelter Island, NY, 2011, Ch. 1

[20] Telgarsky M., Vattani A., Hartigan's Method: k-means Clustering without Voronoi, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)* 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

[21] Scheer D.R., *The Death of Drawing: Architecture in Age of Simulation*, Routledge, UK, 2014, Ch. 1

[22] Maimon O., *Data Mining and Knowledge Discovery Handbook*, Tel-Aviv, Israel, 2005, Chapter 15.