Pesquisa Operacional (2016) 36(3): 575-595
© 2016 Brazilian Operations Research Society
Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope
doi: 10.1590/0101-7438.2016.036.03.0575

DISCOVERING AND LABELLING OF TEMPORAL GRANULARITY PATTERNS IN ELECTRIC POWER DEMAND WITH A BRAZILIAN CASE STUDY

Gabriela Servidone* and Dante Conti

Received March 3, 2016 / Accepted November 7, 2016

ABSTRACT. Clustering is commonly used to group data in order to represent the behaviour of a system as accurately as possible by obtaining patterns and profiles. In this paper, clustering is applied with partitioning-clustering techniques, specifically, Partitioning around Medoids (PAM) to analyse load curves from a city of South-eastern Brazil in São Paulo state. A top-down approach in time granularity is performed to detect and to label profiles which could be affected by seasonal trends and daily/hourly time blocks. Time-granularity patterns are useful to support the improvement of activities related to distribution, transmission and scheduling of energy supply. Results indicated four main patterns which were post-processed in hourly blocks by using shades of grey to help final-user to understand demand thresholds according to the meaning of dark grey, light grey and white colours. A particular and different behaviour of load curve was identified for the studied city if it is compared to the classical behaviour of urban cities.

Keywords: data mining, electricity consumption, load curves, clustering, patterns, time granularity.

1 INTRODUCTION

Getting an optimal balance between generation, distribution and usage of electricity represents a complex problem which involves several interconnected factors: use of natural resources, environmental policies and market evolution. During the last years, electricity market around the world has drastically changed. Electric utilities are facing with new consumer behaviours, government deregulations, dynamic pricing and the establishment of laws and policies focused on developing smart grids and green-smart cities to support sustainability and decreasing environmental impacts [20]. For this reason, utilities are encouraged to reach that balance by constantly trying to adjust energy production to the actual demand [17]. Nowadays, this balance is being achieved at high or top levels, i.e., national or region wide by following energy consumption almost on real-time. At the same time, the support of Information and Communication Technologies (ICT's) and the developing of smart grid systems are generating large amount of data

*Corresponding author.
Departamento de Matemática Aplicada, IMECC, Universidade Estadual de Campinas – UNICAMP, 13083-859 Campinas, Brasil. E-mails: gabi.servidone@gmail.com; conti@ime.unicamp.br

which need to be processed to get non trivial information, so new challenges are emerging within this new context.

Operations Research (OR) and its multidisciplinary approaches are dealing with advanced models to optimize the new scenario inside electricity management. These models include a mixture of classical and emergent techniques supported by Data Mining (DM), Machine Learning (ML), Statistics, Artificial Intelligence (AI), real-time optimization and high computing to manage large amount of data from this complex system [1]. Aggregation is necessary to solve the problem by using top-down approaches to better understand the underlying systems. Under this premise, models are being developed to manage data from top levels, e.g., national, regional demand to down levels that could include neighbourhoods, consumers and details in time-granularity (monthly, daily and hourly blocks). For electricity utilities, the actual demand or the current energy supply in time-granularity is a key factor to manage the new real-time scenario. A load profile, defined as a time-series, describes a graph of the variation of the consumption versus time. This basic definition allows grouping profiles according to customer type, time-granularity, seasonal influences and geographical divisions. The understanding of these profiles is crucial to plan how much electricity is needed to satisfy the demand accordingly with the aggregation level [8, 26].

Clustering in data-driven models is a useful technique to extract patterns and profiles able to better understand and to model systems from a wide variety of fields which include medicine, management, production, finance and resources management as electricity supply and urban water systems [15]. The core is to get knowledge from raw data that can be easy to interpreting. This knowledge is summarized and characterized in meaningful patterns that are usually used as inputs for more complex models oriented to optimize underlying systems and to support decision-making processes. Characterization of load patterns by means of clustering techniques has demonstrated to be useful to produce an accurate description of the behaviour of the demand including levels of time-granularity [23].

This characterization of load patterns is necessary to plan, schedule and design policies and strategies associated to energy generation and distribution. The most important factor is satisfying consumer demand by bidding strategies for market agents (electricity producers) and by getting an efficient balance between energy-supply and consumer demand [15]. As it happens with water demand, electricity demand contains patterns that are influenced by seasonal variables, consumer behaviours and time-granularity [18]. Time-granularity could be seen as an automatic response since data is represented as time-series curves. These curves change in time (monthly, weekly and daily variations until hourly levels). So, the idea to summarize patterns in time-granularity will help to support planning and scheduling in electricity distribution systems. Description of electricity demand based on time-granularity improves the performance of operating and distribution systems and guarantees the long-term planning to control the management between resources and generation of power-supply which are research topics for OR and mathematical models as it was mentioned above. Therefore, these patterns are useful as inputs to forecast new developments on system expansion, design of network distribution and detailing of scheduling of power supply in terms of time blocks and demand. All these factors help in the task to achieve the balance between energy supply and consumer demand, which is the goal to sustainability, profitability and energetic efficiency under the remarkable vision of green and smart management promoted by many countries and societies.

In this research, a data-driven model which uses clustering is described in order to extract representative load patterns obtained from data of a regional electricity utility in a touristic city from South-eastern Brazil. Here, the most outstanding contribution refers to describe profiles from cities whose behaviour in terms of time-granularity demand/electricity supply is hypothetically affected for the own population dynamics over the year due the particular condition of being itself a touristic city or summer resort-town. The study aims to determine patterns from a top-down approach by using clustering according to time-granularity which includes the identification of seasonal groups (top level) until reach daily and hourly patterns (down level) from data. Time-granularity patterns are described and explained to support decisions and policies for a better management of the power supply in this type of non-classical urban cities.

The paper is structured as follows: in Section 2, related works on the field are listed together with Brazilian studies. Section 3 is divided into two subsections. On one hand, theoretical fundamentals about clustering are described, and on the other hand, the methodological approach to solve the problem is presented. Section 4 contains the case study with results. Finally, in Section 5, the conclusions and future works are detailed.

2 RELATED WORKS

From a general point of view, identification and labelling of load patterns in electricity utilities have been studied by applying different techniques and approaches. Many authors adopt classical or emergent techniques, but almost all of them work with data-driven models or model-based reasoning approaches. Some approaches use classical tools from statistics based on time-series analysis. ARIMA and regression models are basically used to deal with fitting curves and forecasting [7, 12, 28]. Other studies try to combine mixed models by applying artificial neural networks with fuzzy logic [2]. Emergent techniques are also studied as cited in [30, 32]. In [32] adaptive and non-parametric models are applied to define patterns and short-term forecast for load profiles in the region of Ontario, Canada. In [30] a weighted nearest neighbours methodology is presented and its performance is compared with others classical statistical techniques. In [9], a research in Sri Lanka applied neural networks to forecast half-hourly electricity load patterns from previous patterns obtained with clustering techniques. In [22] a disaggregated approach for global load curves is described by applying clustering, time-series classical techniques and wavelets.

In a more specific context, delimited by the importance to obtain useful patterns related to how and when consumers use electricity and the connections with time granularity, seasonal influences and clustering or similar techniques, it is possible to list many applications and studies. In [11] a methodology based on the Knowledge Discovery in Databases (KDD) with clustering techniques and classifiers is applied to a case study in Portugal. Here, clustering in two phases is developed to get load patterns (customer differentiation aggregation). The first phase aimed with dimension reduction with Self Organizing Maps (SOM) and then, transformed data is used as input for the clustering engine at the second phase. Results are described not only in terms of time-granularity but also in consumer characterization. Authors in [26] describe the importance of time-granularity for temporal assessment of electricity demand, this temporal aspect of electricity use is significant because electricity is expensive to store, so a scheduling based on the temporal rate of consumption (temporal profiles) guarantees the generation and efficient distribution of the energy. The study used K-means clustering supported by R-software analytics. A similar study in United Kingdom is presented in [8]. Here, profiles are used to optimize the overall energy usage by discovering the amount of overall reduction which occurs during different times (top-down granularity). Similar studies can be found around the world remarking the importance of the clustering approach: [31] refers a case in Korea, [29] reports an overview for load profiles in United Kingdom, [17] presents a study based on geographical aggregation (industrial neighbourhood in central Spain) with top levels in time-granularity. Finally, in [6] a survey is presented to evaluate different techniques based on clustering to group load curves and electricity customer classification.

An interesting real-time approach is presented in [25]; an on-line algorithm is proposed to systematically and efficiently manage the energy consumption data (as data stream) for the optimization of power distribution networks (smart grids).

In the Brazilian context, in [5] a semi-clustering approach based on SOM is used to forecast short-term load patterns (hourly granularity) in a not-mentioned Brazilian electric utility. While in [10] a step by step approach is presented to clustering load curves for an electric utility in the state of Maranhão. No references were found on detecting, extracting and labelling of load patterns similar for our case study, i.e., in a Brazilian context and for a touristic city or summer resort-town.

3 THEORETICAL FUNDAMENTALS AND METHODOLOGICAL APPROACH

Clustering is a process of grouping an unlabelled set of items into a number of clusters such that a similar pattern is associated to every cluster, in other words, clustering divides the data into groups according to the adopted notion of similarity. This data mining (DM) technique can be used in many fields, such as bioinformatics, medicine and marketing. Most of clustering algorithms requires as input the value k (number of clusters) to perform the grouping task. The right value of k is still an open problem. Some researchers try to adapt this value to the problem, experience or final user requirements. However, simulations can be performed in order to obtain k by maximizing quality indexes. In our research, Silhouette [27] and Calinski-Harabasz [21] indexes are used to get this value.

3.1 Clustering: Partitioning Around Medoids (PAM) Algorithm

The PAM algorithm [19], also called as the K-Medoids algorithm, represents a cluster by a medoid.

Initially, the number of desired clusters is an input and a random set of k items is taken to be the set of medoids. After this first step, the algorithm analyses all the non-medoid objects and includes each one of them into one cluster. Thereafter, all medoids are examined to see if an item can represent it better. That is, the algorithm determines whether there is a non-medoid object that should replace one of the existing medoids. In other words, the medoids change their position in each interaction of the algorithm. This process goes until the sum of the dissimilarities between each item and its correspondent medoid is minimized.

PAM algorithm process [3] can be simplified by knowing an input and an output. To start, it is necessary the desired number of clusters, k, and the dataset to classify the containing n items, D. Output returns a set of clusters which minimizes the sum of dissimilarities (*dis*) of all objects to their nearest medoid, as shown below.

$$F(x) = \text{minimize} \sum_{i=1}^{n} \sum_{j=1}^{n} dis(i,\, j) z_{ij}$$

Subject to:

1. $\sum_{i=1}^{n} z_{ij} = 1,\ j = 1, 2, \ldots, n$
2. $z_{ij} \leq y_i,\ i,\ j = 1, 2, \ldots, n$
3. $\sum_{i=1}^{n} y_i = k,\ k = \text{number of clusters}$
4. $y_i, z_{ij} \in \{0, 1\},\ i,\ j = 1, 2, \ldots, n$

where $F(x)$ is the main function to minimize, $dis(i,\, j)$ is the dissimilarity matrix between the points i and j and z_{ij} is the variable which ensures that only dissimilarities with points within the same cluster will be calculated in the main function. And (1) ensures that each point is assigned to only one cluster, (2) ensures that one point is assigned to one medoid which represent a cluster, (3) ensures that there are exactly k clusters and (4) lets the decision variables assume just the values of 0 or 1.

Notice that the standard algorithm aims to minimize the sum of squares, which is equivalent to minimize by Euclidean distance, but other dissimilarity measures or distances could be used as the manhattan distance that are the sum of absolute distances. The algorithm that computes $F(x)$ can be divided in two parts:

Build phase:

1. Choose k entities to become the medoids, or in case these entities were provided use them as the medoids;
2. Calculate the dissimilarity matrix if it was not informed;
3. Assign every entity to its closest medoid;

Swap phase:

1. For each cluster search if any of the entities of the cluster lower the average dissimilarity coefficient, if it does select the entity that lowers this coefficient the most as the medoid for this cluster;
2. If at least one medoid has changed go to (3), else end the algorithm.

As it is described above, PAM algorithm requires the value of k as input. However, R platform supports the "pamk algorithm" [16]. This version of PAM allows performing simulations regarding the value of k. It detects by itself the "best" value of k by maximizing quality measures (Silhouette, Calinski-Harabasz) and returns the clusters without introducing k as input item. The call in software R [24] is:

pamk(data, krange=2:10, criterion="asw", usepam=TRUE, scaling=FALSE,

alpha=0.001, iss=inherits(data, "dist"), critout=FALSE, ns=10, seed=NULL,...)

where the user chooses a *krange* which is the minimum and maximum possible clusters and the criterion can be *"ch"* (Calinski-Harabasz index) or *"asw"* (Silhouette index).

In other words, *pamk* uses for each k the *pam* algorithm and analyses with Calinski-Harabasz and Silhouette indexes which number of clusters is more accurate to represent the database.

3.2 Silhouette

The silhouette index provides a measure of quality of the clusters' separation. Consequently, the larger index will determine which number of cluster is the optimal to the problem. In other words, it validates both the distances between clusters and also between objects inside each cluster. It works as follows [27]: suppose that after the clustering process (by setting the number of clusters as three ($k = 3$)), the objects appear as shown in Figure 1.

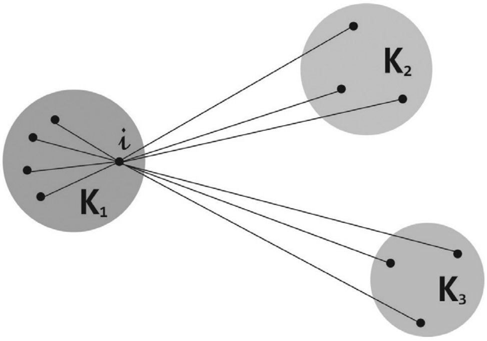


Figure 1 – Clustering distribution of some data ($k = 3$).

So, for each object i in the data set, it is possible to compute $a(i)$ which is the average dissimilarity within-cluster, where the smaller index value implies a better assignment. Afterwards, the dissimilarity between the point i, in this case located in K_1, and all other points in each other cluster is computed and denoted as $dis(i, K_k), \forall k = 2, 3$. When computing $dis(i, K_k)$ for all cluster different from K_1 which is the one that i belongs to, the smallest dissimilarity is chosen and defined as $b(i)$.

Combining $a(i)$ and $b(i)$, silhouette index, $silh(i)$ is defined in equation (1).

$$silh(i) = \frac{a(i) - b(i)}{\max(a(i), b(i))}, \qquad -1 \leq silh(i) \leq 1 \tag{1}$$

However, when the cluster K_1 contains a single object, it is unclear how $a(i)$ should be defined, then, the most correct way is to simplify $silh(i) = 0$.

$Silh(i)$ values closer to 1 refers that the current cluster has a good data classification. The opposite occurs when $silh(i)$ is closer to -1, here it implies that the cluster does not have a good data classification. So, that idea is to maximize the silhouette index. In practice, values (in average) higher than 0.65 are considered as good indicators of clustering.

3.3 Calinski-Harabasz

The Calinski-Harabasz index [21] consists to verify the validity of the cluster in a Euclidean space. This measure is based on the average between cluster and within-cluster. It can be calculated using the following equations:

$$CH(K) = \frac{[trace\mathbf{B}/K - 1]}{[trace\mathbf{W}/n - K]}, \quad \text{for } K \in \mathbb{N} \tag{2}$$

$$trace\mathbf{B} = \sum_{k=1}^{K} n_k \|z_k - z\|^2 \tag{3}$$

$$trace\mathbf{W} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \|x_i - z_k\|^2 \tag{4}$$

where n is the number of points in the dataset. In equation (3), $\mathbf{B}$ denotes the inter-cluster scatter matrix, n_k is the number of points in cluster k and z is the centroid of the entire dataset. In equation (4), $\mathbf{W}$ is the intra-cluster scatter matrix and z_k is the centroid point in cluster k.

For each possible cluster solution, a different Calinski-Harabasz index will take place. The objective is to find the optimal number of clusters which implies the maximum value of the index.

3.4 Methodological Approach

The main objective of this paper is to identify seasonal, daily and hourly patterns inside an initial dataset by following a top-down approach in time granularity. For this purpose, a step by step process is applied as shown in Figure 2.

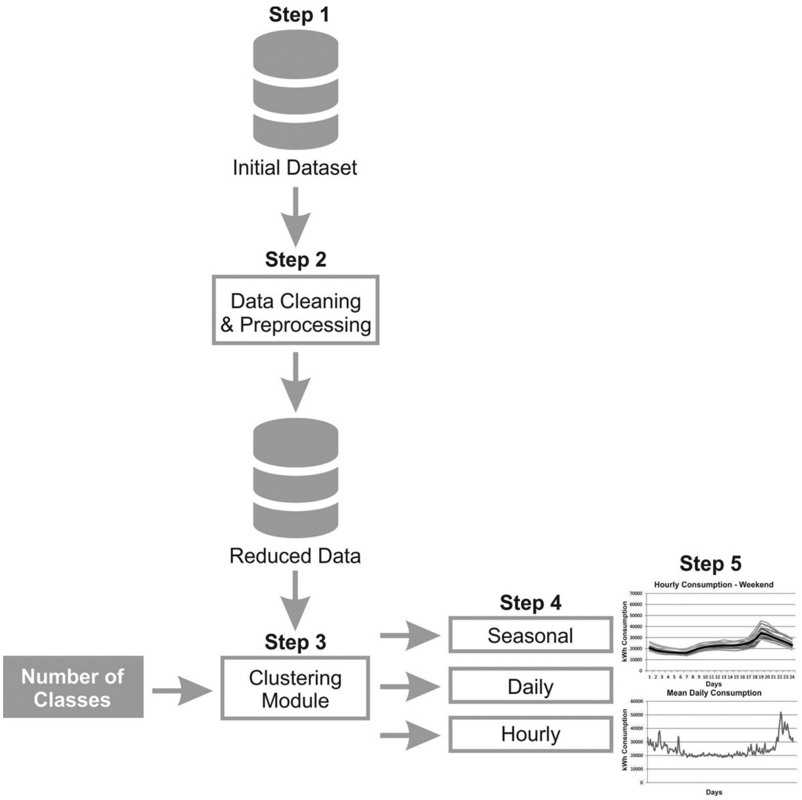


Figure 2 – Step by step flowchart.

Step 1: Initially, the dataset is a numeric table with 365 rows and 24 columns. In other words, the daily electricity consumption information (load curves) is given for each hour of a given year.

Step 2: Pre-processing: here, data is audited to detect erroneous samples and missing values. In addition, datasets are formatted and prepared to be analysed (minable dataset).

Step 3-4: Clustering Process: for each time granularity, the PAM algorithm is applied to detect the clusters and patterns. Firstly, the clustering is performed to the whole dataset returning seasonal patterns, then, the dataset is divided according to these results. From the seasonal major groups, the clustering is separately executed one more time to obtain daily patterns. Finally, transposing the pre-processed tables from daily patterns, clustering is performed to detect hourly patterns.

Step 5: Postprocessing: Here, time blocks are analysed and interpreted. The task aims to summarize all patterns in time-granularity. Labelling is obtained by using shades of grey. Dark grey colour indicates high consumption, white is associated to low consumption and light grey refers average values, respectively.

4 CASE STUDY AND RESULTS

The majority of south-eastern coastal cities in Brazil are economically dependent from tourism. These cities can be labelled as tourist cities or summer resort-towns. Figure 3 shows the population crowd in one of these cities during two different seasons. These cities attract people all around the state for some commonly reasons: inexpensive, compared to other trips available; geographical proximity which increases weekend trips; temperature, agreeable climate in summer. So, the population dynamics is a key factor that can influence over the energy consumption due to this particular situation.

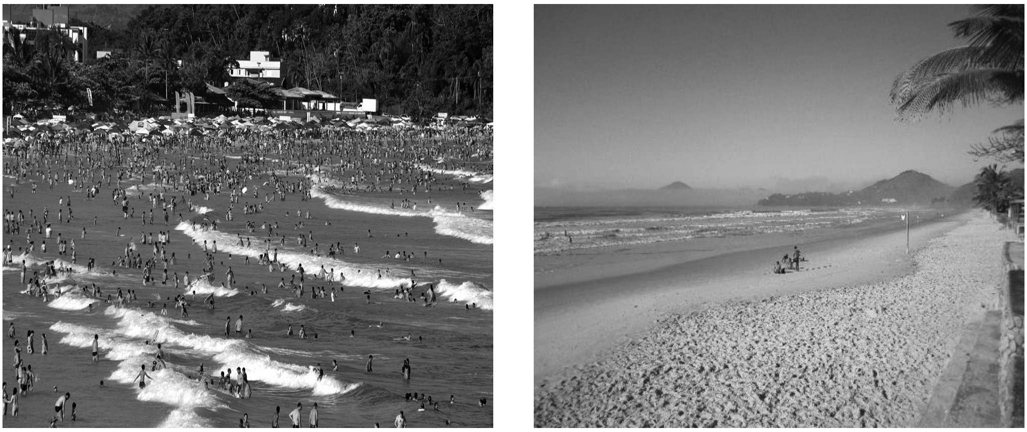
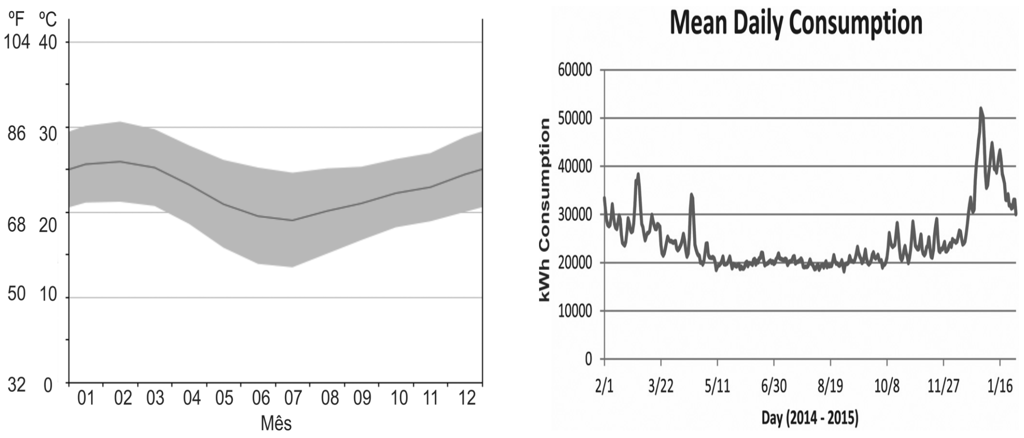


Figure 3 – A beach in a south-eastern coastal city in Brazil. Left during Christmas and new year's eve period. Right during winter season.

In fact, as shown in Figure 4a, the humid coastal tropical climate in these cities has a notably pattern. Between end of November and early March, the mean temperature goes around 25 degrees Celsius, being February the hottest month with an average temperature of 25.5°C, and July, the coldest month with an average temperature of 18.4°C. With Figure 4b, it is easy to see that electrical consumption increases in the exact period when temperature is the highest, sometimes overloading the system. These peaks can cause huge losses and risks for the energy supply.

Under these premises related to climate and population dynamics, our case study was aimed to detect and to label load patterns by considering these possible influences by following a top-down approach in time granularity. Data was collected in 2014 from a local electricity utility, but due to terms of privacy in data, we are not allowed to mention the name of the city, only that it is situated in south-eastern Brazil in the state of São Paulo.

Data was pre-processed and summarized by days (rows) and hours (columns), it contains a set of 365 days from February 1^{st} to January 31^{th}. By using R software and the methodological approach described on Section 3.4, results are listed as follows:



(a) Mean, highest and lowest temperature by month. (http://pt.climate-data.org/location/34847/).

(b) Mean consumption starting in 1^{st} February (2014) and ending in 30^{th} January (2015). Note that the highest peak is in January 1^{st}.

Figure 4 – Figures (a) and (b) indicate mean temperature and energy consumption for the case study.

RESULTS

Pamk was used to perform clustering. Obtained $k's$ were analysed by comparing them with other experiences (related works) and researches. Relationships between obtained clusters and the top-down approach in time granularity are detailed to identify and to label the representative patterns for the load curves. By this way, results are divided into: seasonal influences, daily patterns and hourly patterns, respectively.

Clustering Results: Seasonal and daily patterns

Firstly considering the whole dataset, an initial clustering was performed. The k with the best quality indexes was $k = 2$. From a logical point of view according to the underlying system, this k value seems to be accurate. Note that patterns in seasonal influences commonly take values from 2 to 4: 2 for Summer/Spring and Winter/Autumn influences, 3 for seasonal influences with transitional months or 4 for the complete seasons (Summer, Autumn, Winter and Spring) separately, as it commonly happens in classical urban cities. Power demand in common urban cities is influenced by seasons. The most likely patterns are associated to Summer/Spring and Winter/Autumn days in countries in which have the four distinct seasons and subtropical climate as Brazil. In these cities, power demand is also influenced by the typology of the days, i.e., working days (week days from Monday to Friday) and weekend/holidays (Saturday, Sunday and holidays). Many authors refer this as the classical load characterization which contain four manly patterns: (1) Summer/Spring, working days, (2) Summer/Spring, weekends and holidays, (3) Winter/Autumn, working days and (4) Winter/Autumn, weekends and holidays [9, 17, 20, 30, 31].

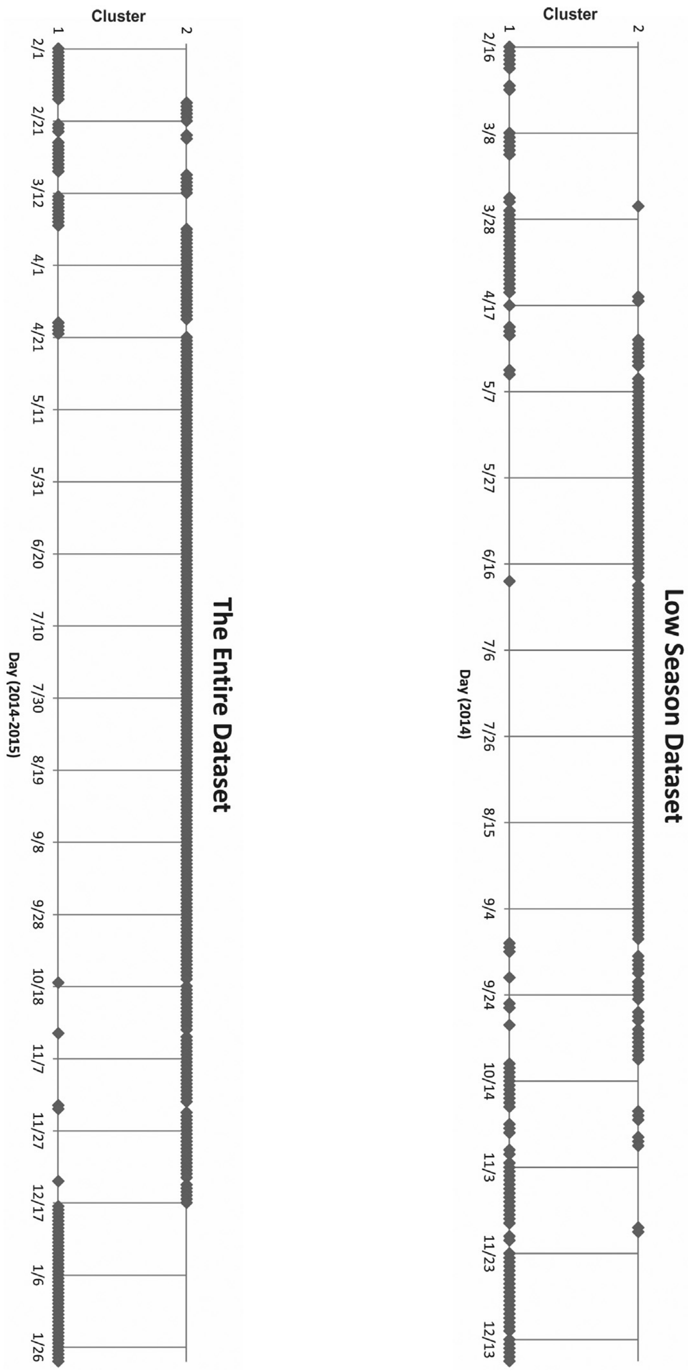
In our case of study, a seasonal influence was detected, but a little more different than the classical classification of the urban cities mentioned before. Figure 5(a) shows the year 2014 classified into 2 clusters. Days labelled as 1 represent days with high consumption and days with label 2 describe low consumption patterns. In other words, seasonal influences are clearly differentiated, and do not follow the classical classification: summer season and some important holidays for Brazilian culture, i.e., high pattern: from December 19 to February 15, Carnival and Easter days; and low pattern: the rest of the days belong to Spring, Winter and Autumn seasons. Inside these two clusters no daily typology labelled as “working day (from Monday to Friday)” or “weekend and holidays” was found as usually happens in classical urban cities. This is a notorious behaviour that could describe special situations related to tourist city or summer resort-towns, because in urban cities is common to find a clear typology of days. In order to validate this behaviour, clustering was separately performed for these two blocks during the year, but no daily patterns were found for the summer block, but something interesting occurred during the block containing the low pattern consumption (rest of the year).

In Figure 5(b) (low pattern consumption) it is possible to note that inside this block, a daily typology could be inferred, even if this pattern is not present during the whole block. If (in Fig. 5(b)) the days between 184 to 231 (September 12 to October 29) were amplified (see Fig. 6), a daily pattern inside the low season block is clearly distinguished, i.e., a type of division in working days and weekends.

A meticulous analysis (by performing a new clustering over this subset of the low pattern block) revealed that most of these days were indeed separated as it follows: from Monday to Thursday as one cluster (similar to the classical classification label of “working day” in urban cities) and a second cluster which includes days from Friday to Sunday (weekends and holidays). However, not all of them were correctly grouped; some of the working days were classified as if they were weekends or holidays. Even so, the clustering was able to perform a good work since that some of the misclassified working days belong to holidays in the Brazilian calendar. For example the days from October 13^{th} to 16^{th}, even being working days, they take part of the pre and post holiday of October 12^{th} in Brazil. Figure 8 details the final daily typology.

This particular period of 48 days from September to October contains 21 days labelled as weekends, considering Fridays, and 27 working days. Only six of the week days were not correctly classified, hence, the error in working days is 22% (6 days out of 27). On the other hand, removing those 4 misclassified days, the error goes down to 7%. With the same approach, there were six holidays improperly grouped. Consequently, the average error for the weekends is 28% (6 days out of 21). Removing the 3 wrongly cluster days, the error decreases to 14%. Thus, the total error is 10.41% (5 days out of 48) for the whole period, so clustering results offer an accuracy of almost 90%.

After these clustering processes (seasonal influences and daily typology), the total load curve for that city over the year can be divided into 4 major groups or patterns: (1) High Season (high consumption-majority of summer days), (2) Low season (rest of the year, excepting the 48 days mentioned for September and October), a transitional pattern with daily typology (September



(a) Represents clustering of the whole year.

(b) Represents clustering of low season days.

Figure 5 – Distribution of days belonging to 2014.

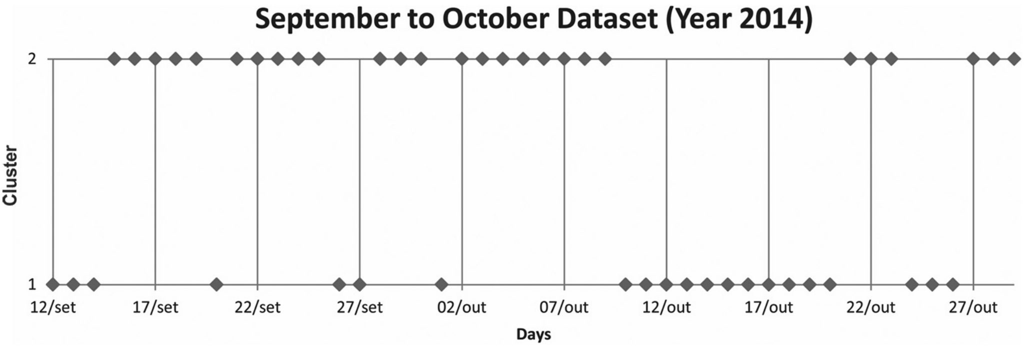


Figure 6 – Amplification window for the 48 days from Low Season Dataset.

and October): (3) a Lower consumption pattern – Working/week day label and (4) Higher consumption – Weekend and holidays. Figure 7 shows the load curves by days of these profiles and Figure 9 describes the distribution of these 4 patterns all year long.

Clustering results: Hourly patterns and interpretation

Once seasonal and daily patterns were found, the top-down framework was focused on detecting more detailed information to support hourly scheduling for electric utilities. In this case, clustering with PAM algorithm is performed over the 4 databases generated from the whole dataset which was divided according to the 4 patterns obtained in the previous experiments.

In Brazil, there is a policy which consists to delay one hour from October to February in order to save electricity with day light. For this reason and before performing the hourly clustering, all databases were adjusted to have the same solar hour.

Hourly clustering is described in Figure 10. In high season, there are only two different hourly blocks while in the rest of the blocks, three different hourly blocks were obtained. Here, an analogy with the colours of the TLP is introduced to better interpreting of the results. This analogy which takes part of a more complex post-processing tool [13] has demonstrated to be useful to overcome the gap between clustering results and decision-making. TLP constructed over conditional means or medians has been applied with good results in different fields [4, 14]. Based on this, Figure 10 shows levels of hourly consumption according to local means and its analogy with TLP colours: High (dark grey which refers red color within the TLP), Medium (light grey, analogy of yellow color of the TLP) and Low (white, analogy with the green color of the TLP).

It can be clearly seen that between 6 pm to 10 pm there is a huge consumption of electricity. Exactly the time in which the majority of people are returning home, turning on the TV, taking a shower, and it is also the moment when sunlight disappears and the necessity of electricity becomes constant. Analogously easy to be interpreted, the hours between 1 am and 8 am are the sleeping hours, consequently the consumption decreases, except in high season days where night activities seems to change the ordinary behaviour.

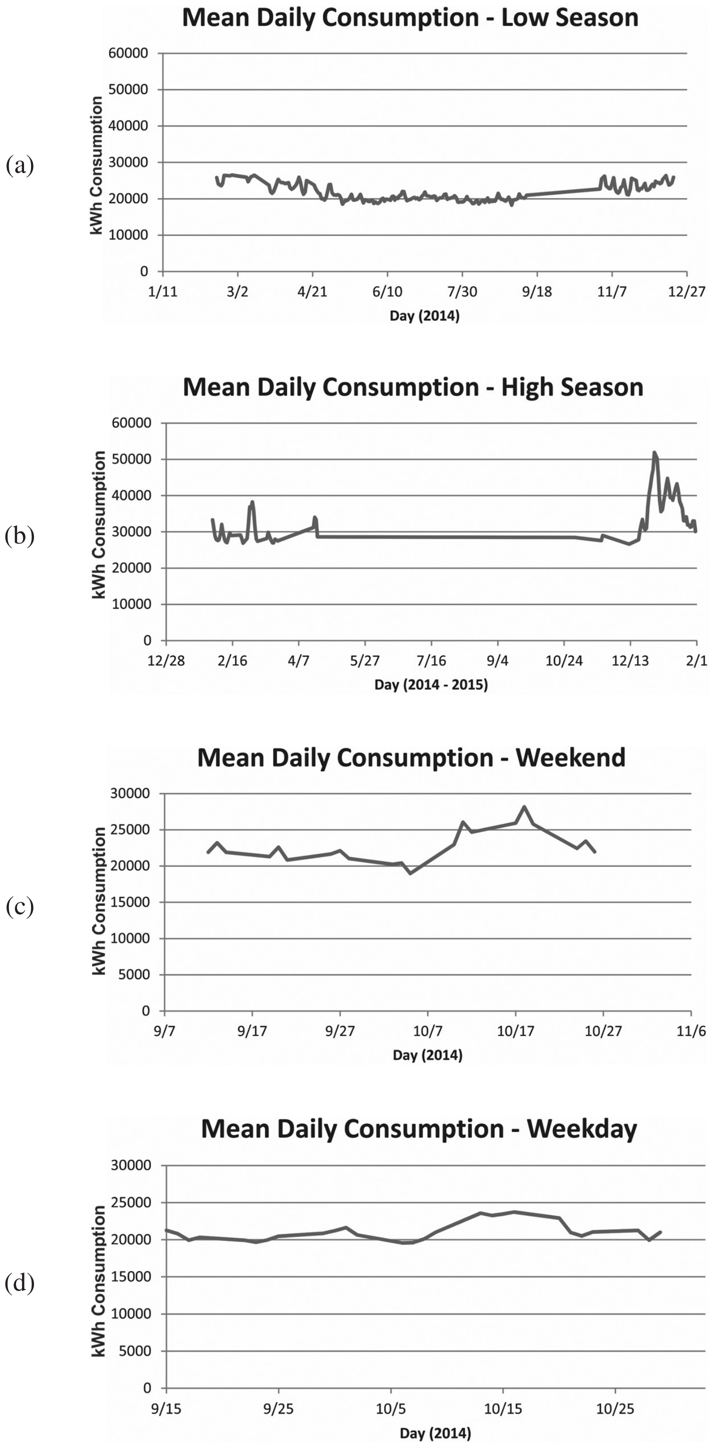
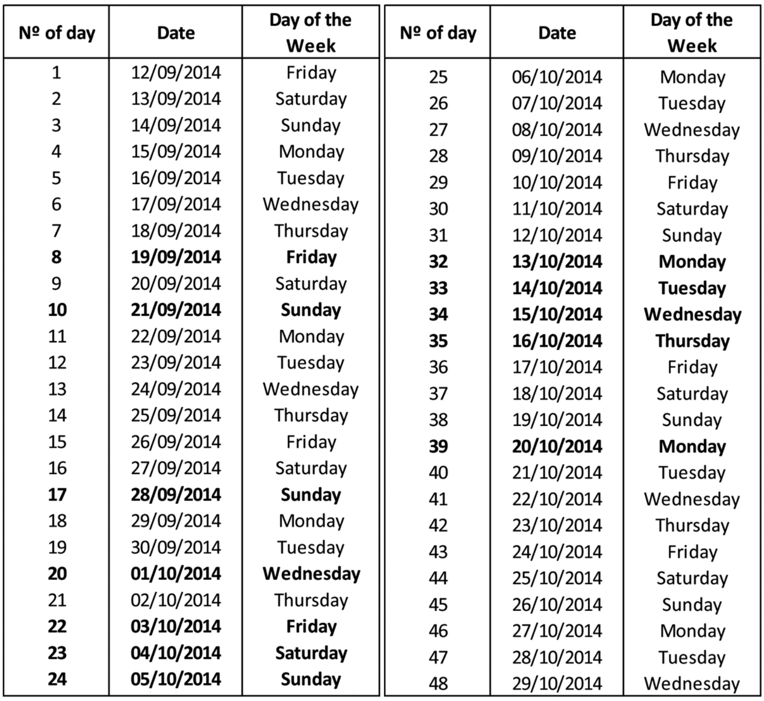


Figure 7 – Figures (a) to (d) – Average daily power demand in the four patterns of the case study.



Nº of day	Date	Day of the Week
1	12/09/2014	Friday
2	13/09/2014	Saturday
3	14/09/2014	Sunday
4	15/09/2014	Monday
5	16/09/2014	Tuesday
6	17/09/2014	Wednesday
7	18/09/2014	Thursday
8	**19/09/2014**	**Friday**
9	20/09/2014	Saturday
10	**21/09/2014**	**Sunday**
11	22/09/2014	Monday
12	23/09/2014	Tuesday
13	24/09/2014	Wednesday
14	25/09/2014	Thursday
15	26/09/2014	Friday
16	27/09/2014	Saturday
17	**28/09/2014**	**Sunday**
18	29/09/2014	Monday
19	30/09/2014	Tuesday
20	**01/10/2014**	**Wednesday**
21	02/10/2014	Thursday
22	**03/10/2014**	**Friday**
23	**04/10/2014**	**Saturday**
24	**05/10/2014**	**Sunday**
25	06/10/2014	Monday
26	07/10/2014	Tuesday
27	08/10/2014	Wednesday
28	09/10/2014	Thursday
29	10/10/2014	Friday
30	11/10/2014	Saturday
31	12/10/2014	Sunday
32	**13/10/2014**	**Monday**
33	**14/10/2014**	**Tuesday**
34	**15/10/2014**	**Wednesday**
35	**16/10/2014**	**Thursday**
36	17/10/2014	Friday
37	18/10/2014	Saturday
38	19/10/2014	Sunday
39	**20/10/2014**	**Monday**
40	21/10/2014	Tuesday
41	22/10/2014	Wednesday
42	23/10/2014	Thursday
43	24/10/2014	Friday
44	25/10/2014	Saturday
45	26/10/2014	Sunday
46	27/10/2014	Monday
47	28/10/2014	Tuesday
48	29/10/2014	Wednesday

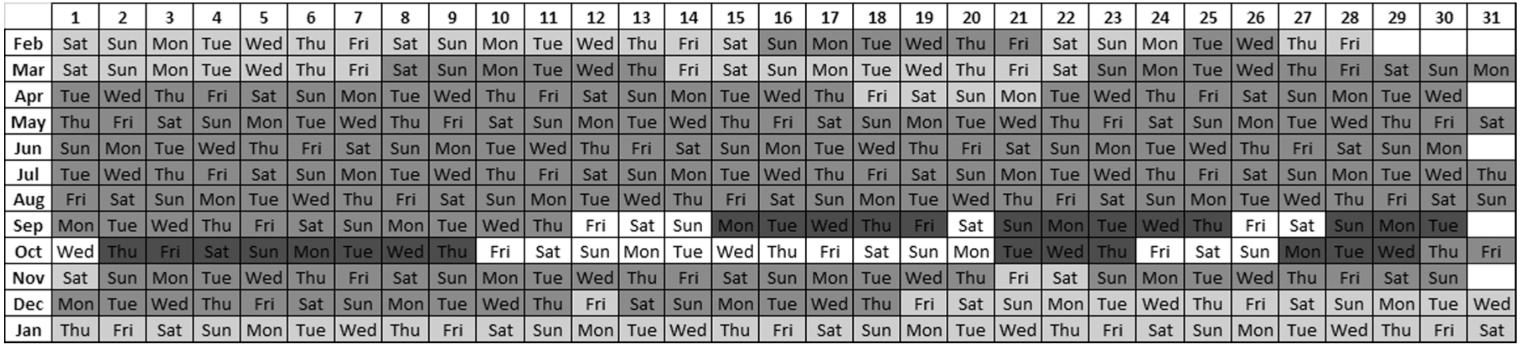
Figure 8 – Classification of each day in working days and weekends, bold indicates which day was wrongly classified (Year 2014).

Finally, Figure 11 shows the shape of the load curves for each pattern during the hourly block, here every line represents a day and the black one represents the centre of the cluster as the mean of the values belonging each cluster.

5 CONCLUSIONS AND FUTURE WORKS

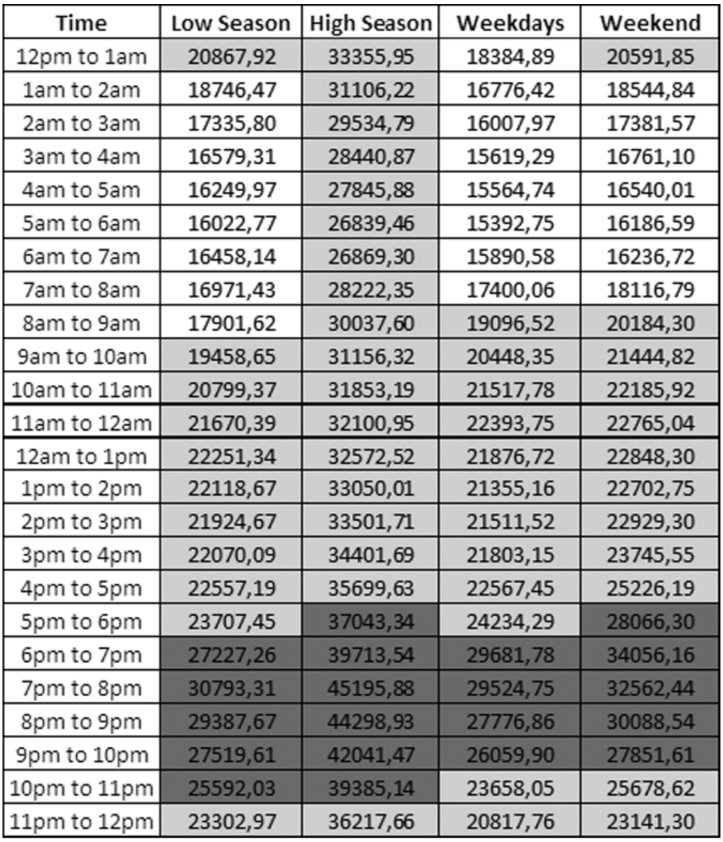
A top-down approach in time granularity with partitioning clustering techniques (PAM) was applied to detect, label and describe representative patterns associated to load curve (demand/energy comsumption) from a south-eastern coastal city in Brazil. Load curves and its typification is a key factor to better support electric utilities to achieve actual demand and to optimize energy scheduling and consumption forecasting.

This research proposed a data-driven approach which allowed creating average seasonal curves in hourly granularity for 4 different patterns differentiated by seasonal and daily typology influences. The case study related to a tourist city or summer resort-town described particular behaviours that are quite different to the classical behaviours of urban cities. In this case, load curves and typification are essencial to predict and plan certain action plans for holidays by example, which allows to prevent a system overload. Clustering performance in accuracy reached levels of 90% approximately.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Feb	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri			
Mar	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon
Apr	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	
May	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Jun	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	
Jul	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu
Aug	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Sep	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	
Oct	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri
Nov	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	
Dec	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed
Jan	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat

Figure 9 – Distribution of the identified patterns. The colors are represented from white to dark grey in this order: weekends, high season, low season and working days.



Time	Low Season	High Season	Weekdays	Weekend
12pm to 1am	20867,92	33355,95	18384,89	20591,85
1am to 2am	18746,47	31106,22	16776,42	18544,84
2am to 3am	17335,80	29534,79	16007,97	17381,57
3am to 4am	16579,31	28440,87	15619,29	16761,10
4am to 5am	16249,97	27845,88	15564,74	16540,01
5am to 6am	16022,77	26839,46	15392,75	16186,59
6am to 7am	16458,14	26869,30	15890,58	16236,72
7am to 8am	16971,43	28222,35	17400,06	18116,79
8am to 9am	17901,62	30037,60	19096,52	20184,30
9am to 10am	19458,65	31156,32	20448,35	21444,82
10am to 11am	20799,37	31853,19	21517,78	22185,92
11am to 12am	21670,39	32100,95	22393,75	22765,04
12am to 1pm	22251,34	32572,52	21876,72	22848,30
1pm to 2pm	22118,67	33050,01	21355,16	22702,75
2pm to 3pm	21924,67	33501,71	21511,52	22929,30
3pm to 4pm	22070,09	34401,69	21803,15	23745,55
4pm to 5pm	22557,19	35699,63	22567,45	25226,19
5pm to 6pm	23707,45	37043,34	24234,29	28066,30
6pm to 7pm	27227,26	39713,54	29681,78	34056,16
7pm to 8pm	30793,31	45195,88	29524,75	32562,44
8pm to 9pm	29387,67	44298,93	27776,86	30088,54
9pm to 10pm	27519,61	42041,47	26059,90	27851,61
10pm to 11pm	25592,03	39385,14	23658,05	25678,62
11pm to 12pm	23302,97	36217,66	20817,76	23141,30

Figure 10 – Labelling of hourly patterns by using shades of grey: dark grey to high, light grey to medium and white to low consumption.

Patterns in this type of cities seem to be more affected by population dynamics than the typical consumer behaviours found in urban cities. Seasonal influences were labelled in two major groups. On one hand, summer days have load curves with high consumption and for this, careful policies on energy distribution should be planned in order to avoid lacks on demand. Hourly blocks for summer days also suggested high and medium levels in energy demand, so hourly scheduling remains in warning levels during these days. On the other hand, low season days presented a consumption pattern more homogeneous and correlated with the classical consumer behaviour (three typical and hourly blocks during the day). However, this research was able to detect transitional months (September and October) where the daily typology is also feasible and different. This detailed information is useful to a better scheduling of energy distribution. Most of the days during the year are in a "safe" demand, but for September and October a more accurate scheduling can be performed due to the differences between week days (Monday to Thursday) and weekends/holidays (Friday to Sunday) which could improve "savings" in the energy supply planning.

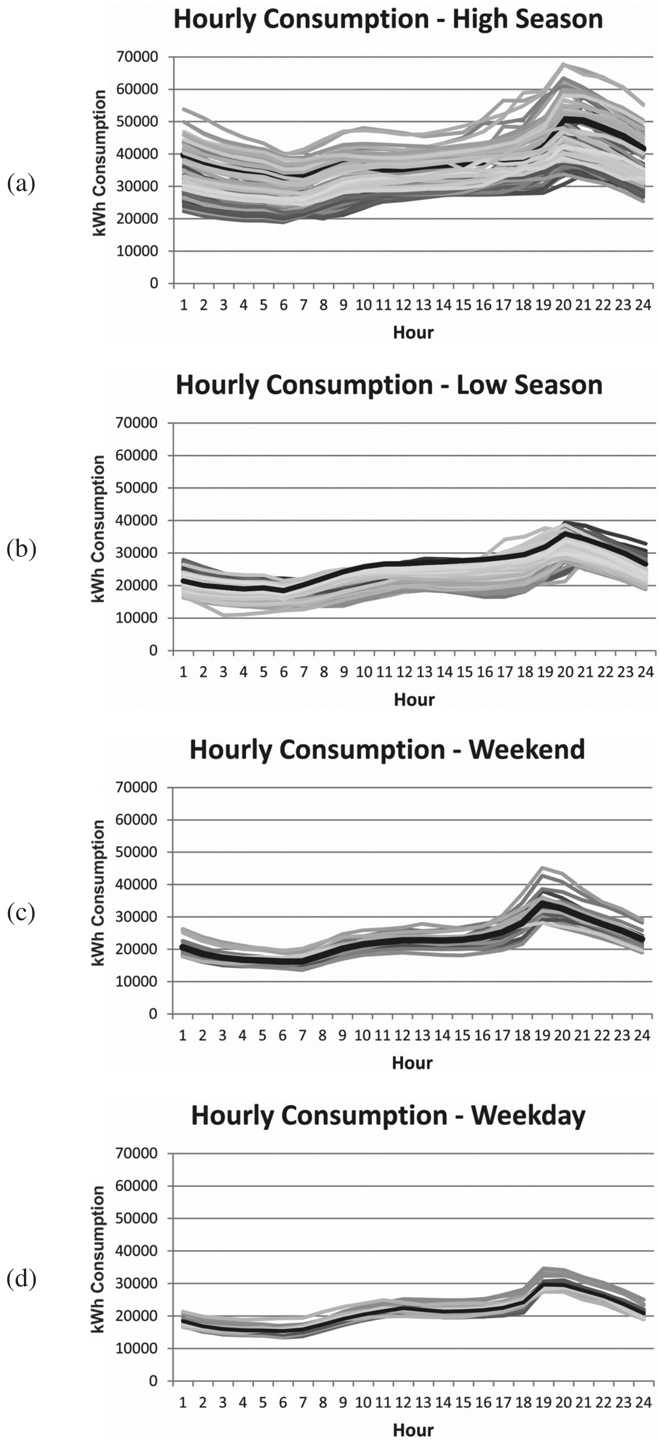


Figure 11 – Figures (a) to (d) – Shape of hourly load curves for the 4 patterns in the case study.

The patterns obtained in this research are being used in supervised algorithms aimed to forecast hourly consumption on real-time. The four main patterns and their hourly distribution will be used to training Support vector machine regression models. Data will be divided into two subsets: training and test set, respectively for each labelled pattern. Then, forecasting models will be developed for the future hourly demand given the current hourly demand. Thus, previsions could be obtained on real time only with the actual data collected from real-time systems (SCADA) conditioned by inputs already labelled within the four patterns. In other words, given a day, this day is labelled according to its season and typology, then, regressors are calculated by considering the obtained profiles in hourly level. Some experiments are being developed in this field with good results in urban water systems [4]. More disaggregated studies can be also performed to achieve profiles for individual customers or neighbourhoods inside that city. Here, smart grid models supported by smart meter (electronic devices which record consumption of electric energy in intervals of an hour or less) jointly with clustering approaches are being analysed.

REFERENCES

[1] AHAT M, AMOR M, BUI A, BUI G, GUÉRARD G & PETERMANN C. 2013. Smart Grid and Optimization. *American Journal of Operations Research*, **3**(1): 196–206.

[2] AMJADY N. 2006. Day-Ahead Price Forecasting of Electricity Markets by a New Fuzzy Neural Network. *IEEE Transactions on Power Systems*, **21**(2): 887–896.

[3] BHAT A. 2014. K-Medoids Clustering Using Partitioning Around Medoids for Performing Face Recognition. *International Journal of Soft Computing, Mathematics and Control*, **3**(3): 1–12.

[4] CANDELIERI A, CONTI D, CAPELLINI D & ARCHETTI F. 2014. Urban Water Demand Characterization and Short-term forecasting – the ICEWater Approach. Proceedings of the 2014 International Conference on Hydroinformatics, Paper 250, New York, USA.

[5] CARPINTEIRO OAS & REIS AJR. 2005. A SOM-based Hierarchical Model to Short-Term Load Forecasting. *IEEE Power Tech*, 1–6.

[6] CHICCO G, NAPOLI R & PIGLIONE F. 2006. Comparisons Among Clustering Techniques for Electricity Customer Classification. *IEEE Transactions on Power Systems*, **21**(2): 933–940.

[7] CONEJO AJ, PLAZAS MA, ESPÍNDOLA R & MOLINA AB. 2005. Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models. *IEEE Transactions on Power Systems*, **20**(1): 1035–1042.

[8] DENT I, AICKELIN U & RODDEN T. 2011. The Application of a Data Mining Framework to Energy Usage Profiling in Domestic Residences Using UK Data. Buildings Don't Use Energy, People Do: Research Students' Conference, 28 June 2011, Bath, England. (Unpublished).

[9] DESHANI KAD, ATTYGALLE MDT, HANSEN LL & KARUNARATNE A. 2014. An Exploratory Analysis on Half-Hourly Electricity Patterns Leading to Higher Performances in Neural Network Predictions. *International Journal of Artificial Intelligence & Applications*, **5**(3): 37–51.

[10] FERREIRA AMS, CAVALCANTE CAMT, FONTES CHO & MARAMBIO JES. 2013. Pattern Recognition of Load Profiles in Managing Electricity Distribution. *International Journal of Industrial Engineering and Management*, **4**(3): 117–122.

[11] Figueredo V, Rodrigues F, Vale Z & Gouveia JB. 2005. An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques. *IEEE Transactions on Power Systems*, **20**(2): 596–602.

[12] García RC, Contreras J, van Akkeren M & García JBC. 2005. A GARCH Forecasting Model to Predict Day-Ahead Electricity Prices. *IEEE Transactions on Power Systems*, **20**(2): 867–874.

[13] Gibert K et al. 2008. Response to TBI-neurorehabilitation through an AI & Stats hybrid KDD methodology. *Medical Archives*, **62**(3): 132–135.

[14] Gibert K, Conti D & Sánchez-Marré M. 2012. Decreasing Uncertainty When Interpreting Profiles through the Traffic Lights Panel. *Advances in Computational Intelligence – Communications in Computer and Information Science*, **298**(1): 137–148.

[15] Han J, Kamber M & Pei J. 2011. Data Mining: Concepts and Techniques, 3rd edition.

[16] Hennig C. 2015. fpc: Flexible Procedures for Clustering. R package version 2.1-10. http://CRAN.R-project.org/package=fpc.

[17] Hernández L, Baladrón C, Aguiar JM, Carro B & Sánchez-Esguevillas A. 2012. Classification and Clustering of Electricity Demand Patterns in Industrial Parks. *Energies*, **5**: 5215–5228.

[18] Herrera M, Iziquierdo LTJ & Perez-Garcia R. 2010. Predictive models for forecasting hourly urban water demand. *Journal of Hydrology*, **387**: 141–150.

[19] Ibrahim LF & Harbi MHA. 2012. Using Modified Partitioning Around Medoids Clustering Technique in Mobile Network Planning. *International Journal of Computer Science Issues*, **9**(6): 299–308.

[20] Martinez-Álvarez F, Troncoso A, Riquelme JC, Aguilar-Ruiz JS & Petermann C. 2011. Energy Time Series Forecasting Based on Pattern Sequence Similarity. *IEEE Transactions on Knowledge and Data Engineering*, **23**(8): 1230–1243.

[21] Maulik U & Bandyopadhyay S. 2002. Performance Evaluation of Some Clustering Algorithms and Validity Indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(12): 1650–1654.

[22] Misiti M, Misiti Y, Oppenheim G & Poggi JM. 2010. Optimized Clusters for Disaggregated Electricity Load Forecasting. *Statistical Journal*, **8**(2): 105–124.

[23] Plaza MA, Conejo AJ & Prieto FJ. 2005. Multimarket Optimal Bidding for a Power Producer. *IEEE Transactions on Power Systems*, **20**(4): 2041–2050.

[24] R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.jstatsoft.org/v65/i06.

[25] Ranjan P, Charalampos C, Frincu M & Prasanna V. 2015. On Online Time Series Clustering for Demand Response: Optic – A Theory to Break the ‘Curse of Dimensionality’. *ACM Sixth International Conference on Future Energy Systems*, 95–100.

[26] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF & Webber ME. 2014. Clustering Analysis of Residential Electricity Demand Profiles. *Applied Energy*, **135**: 461–471.

[27] Rousseuw PJ. 1987. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**: 53–65.

[28] SUMER KK, GOKTAS O & HEPSAG A. 2009. The Application of Seasonal Latent Variable in Forecasting Electricity Demand as an Alternative Method. *Energy Policy*, **37**: 1317–1322.

[29] THUMIM J. 2014. Investigating the potential impacts of Time of Use (TOU) tariffs on domestic electricity customers. Technical report – Centre for Sustainable Energy, Bristol (UK).

[30] TRONCOSO A, RIQUELME JM, EXPÓSITO AG, RAMOS JLM & RIQUELME JC. 2006. Electricity Market Price Forecasting Based on Weighted Nearest Neighbors Techniques. *IEEE Transactions on Power Systems*, **22**(3), 1294–1301.

[31] YU IH, LEE JK, KO JM & KIM SI. 2005. A method for classification of electricity demands using load profile data. *Computer and Information Science*, **4**: 164–168.

[32] ZERIPOUR H, BHATTACHARYA K & CANIZARES CA. 2006. Forecasting the Hourly Ontario Energy Price by Multivariate Adaptive Regression Splines. *IEEE Power Engineering Society General Meeting*, Montreal, Que.