International Congress on Environmental Modelling and Software

9th International Congress on Environmental Modelling and Software - Ft. Collins, Colorado, USA - June 2018

Jun 27th, 9:00 AM - 10:20 AM

Characterization of electric energy consumption with clustering techniques: a case study in Northern Mexico

Dante Conti
Mathematical Science Center Applied to Industry CEPID-CeMEAI, University of Sao Paulo and State University of Campinas, Brazil, danteconti73@gmail.com

Karina Gibert
Dep. Statistics and Operations Research; Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence Research Center; Research Institute of Science and Technology for Sustainability; Universitat Politècnica de Catalunya-BarcelonaTech, Spain, karina.gibert@upc.edu

Daniela De La Rosa
RICH IT, Mexico City - Mexico, ddelarosa@richit.ai

iEMSs International Congress on Environmental Modelling and Software

9th International Congress on Environmental Modelling and Software
Fort Collins, Colorado, USA, Mazdak Arabi, Olaf David, Jack Carlson, Daniel P. Ames (Eds.)
https://scholarsarchive.byu.edu/iemssconference/2018/

Characterization of electric energy consumption with clustering techniques: a case study in Northern Mexico

Conti Dante[1], Gibert Karina[2], De La Rosa Daniela[3]
[1] Mathematical Science Center Applied to Industry CEPID-CeMEAI, University of Sao Paulo and State University of Campinas, Brazil, conti@ime.unicamp.br
[2] Dep. Statistics and Operations Research; Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence Research Center; Research Institute of Science and Technology for Sustainability; Universitat Politècnica de Catalunya-BarcelonaTech, Spain), karina.gibert@upc.edu
[3] Rich IT, Mexico City – Mexico, ddelarosa@richit.ai

Abstract: Nowadays, for electricity operators it is crucial to get a balance between generation & distribution of energy versus current demand, so a better characterization of load profiles would help operators to achieve that balance. Data mining techniques are frequently used to discover patterns of energy consumption in time-granularity. This consumption is affected for seasonal trends, weather, type of days and hourly blocks. This research is aimed to determine differentiated profiles by applying clustering techniques to divide data in groups labeled in relation with top-down levels in time-granularity: seasons (top level), typology of days and hourly blocks (down level). A database containing energy consumption and weather variables measured daily in hourly blocks from January 2004 until August 2017 in Northern Mexico was used during the experiments. Clustering results determined that energy consumption all year long can be characterized in five load profiles according to seasons and daily blocks: Summer/Working days (Mondays to Fridays), Summer/Saturdays, Summer/Sundays and Holidays, Rest of the year/Mondays to Saturdays and Rest of the year/Sundays and Holidays. A final grouping task in hourly granularity was developed within the five temporal profiles separately and post-processed by using the Traffic Light Panel (TLP) tool to help final-user to understand hourly demand thresholds according to the meaning of the colors of the TLP's, respectively. Finally, results were validated with experts on the field and some works are now focused on the implementation of an API with short-term forecasting tools, by considering the load profiling discovered so far.

Keywords*:* clustering; electricity consumption; load curves; patterns; time granularity.

1 INTRODUCTION

During the last years, electricity operators all over the world face with new challenges related to electricity market: new consumer behaviors, law changes, deregulation, dynamic pricing and trends which are focused on developing smart grids and green-smart cities for sustainability and reducing of environmental impacts when generating and distributing energy (Martinez-Alvarez et al., 2011). For this reason, operators need to get an efficient balance between production and actual demand, almost in real-time, so forecasting techniques together with a detailed description of consumption patterns are crucial to optimize activities regarding production, scheduling, and load distribution in energy systems (Hernandez et al., 2012).

A load profile, defined as a time-series, describes the variation of the energy consumption or power demand versus time. As it happens in urban water distribution systems, the power demand is also affected by seasonal trends, weather influences and the variability of consumption in time-granularity, especially on type of days and hours (Candelieri, 2017). The understanding and description of consumption patterns denoted here as characterization of profiles, is a key factor to plan how much electricity is needed to satisfy the demand according to different aggregation levels in time blocks: months, weeks, days and hours (Deshani et al., 2014) (Rhodes et al., 2014).

Clustering is a useful technique which groups data or objects into clusters, represented as vectors in a multi-dimensional space, such that some similarity/distance measure is maximized/minimized within groups and minimized/maximized between groups (Han et al., 2011). Results of clustering are then summarized into patterns or profiles able to better understand and to model systems from a wide variety of fields including energy consumption (Plaza et al., 2005). This knowledge is commonly used as inputs for more complex models or to support decision-making processes.

Characterization of electric energy consumption by means of clustering techniques has demonstrated to be productive in the determination of an accurate description of the demand behavior in different levels of time-blocks (Servidone & Conti, 2016). In this work, clustering techniques, specifically Hierarchical Clustering is applied to historical data of load curves to obtain profiles of consumption in different levels of time-granularity as an input tool that will feed a forecasting engine to be developed in a future work. Also, obtained profiles will be post-processed by means of the Traffic Light Panel tool, TLP (Gibert et al., 2008) to assist end-user in the interpretation of the results.

Under this premise, the main objective of this research is to execute clustering over the historical load curves to extract and characterize differentiated profiles into levels of time-blocks according to a top-down approach which begins with seasonal divisions, then daily patterns and finally, hourly patterns inside the dataset. From this division, the connection between belonging to a certain cluster and the model of prediction for each cluster could be found to customize the forecasting engine. Therefore, the prediction would be generated by means of the corresponding model associated for each differentiated profile (cluster). This prediction engine will act in response, almost in real-time (each hour), by returning the energy consumption to be scheduled each day (in hours blocks) with data registered by the primaries hours of the current day which is received and transmitted from SCADA systems.

The paper has the following structure: section 2 describes the methodological approach. Section 3 introduces the application domain. Section 4 shows experimental results. Section 5 provides discussion about the time-blocks profiles in terms of the energy characterization and their corresponding TLPs, and finally, section 6 details conclusions and future works.

2 METHODOLOGICAL APPROACH

The methodological approach follows the steps of the KDD process (Knowledge Discovery in Databases) introduced by Fayyad (1996). The Data Mining engine is based on clustering (hierarchical clustering) and postprocessing step uses the Traffic Light Panel - TLP (Gibert et al., 2008) to present results to the end-user and to support the conceptualization of the clusters.

From a general point of view, the step by step approach is:

Step 1: Determine relevant data, collect and prepare for the analysis. Perform some exploratory data analysis to get some previous insights from data.
Step 2: Perform clustering over the dataset. Repeat clustering by sub-setting the dataset according to the top/down levels in time-blocks: months (seasons), days and finally, hours. Results are grouped in classes or profiles.
Step 3: Obtain basic statistics per class to label representative centroids of the clusters.
Step 4: Postprocessing: Get the polarity of the variables according to their meaning (semantics) from experts and perform color-association. Build the TLP.
Step 5: Show the TLP to the experts and ask them to conceptualize the classes by providing class labels (knowledge production step).
Step 6: Associate decisions to every class together with experts (if applicable).

The approach was developed to be:
(a) Completely data-driven; it considers as input the historical energy consumption data;
(b) Enriched with external data associated to weather conditions and calendar of special dates in Mexico obtained from open data sources;
(c) based on a multiple-stage in time-granularity (top/down levels): (i) identifying and characterizing of seasonal and daily consumption patterns and (ii) grouping, labeling and postprocessing of (i) in hourly blocks by means of the TLP. So, the approach deals with nonlinear variability of energy consumption in different time-blocks, automatically characterizing behavior-related differences of different types of days and hours of the day including seasonal influences;
(d) Useful as an input to customize a short-term forecasting engine to be developed in a future research.

Some details of the approach are listed as follows:

-Hierarchical clustering is applied by using the Ward´s method (Ward, 1963) and cosine as the similarity measure to perform the clustering task. A previous analysis to select the similarity measure was done by considering similarity in time, in shape and in change of the load curves. Cosine similarity was chosen due to good results in previous works developed by the authors in similar fields (Candelieri & Archetti, 2014) (Servidone & Conti, 2016). Selection of the number of clusters “K” is supported by quality index, in this research, Calinski-Harabasz (1974) is used to confirm this number that will be used to cut the dendrogram obtained from the clustering process.

- Once the clustering procedure in the different levels (time-granularity) finished, the centroid of each cluster is selected to represent the energy consumption pattern for the set of time-series belonging to that cluster. When using the top/down time-granularity levels embedded clustering processes are performed at level of time-granularity and results produce subsequent subdivisions of higher level clusters. Thus, K_1 clusters and consequently, K_1 subsets of data are obtained when clustering is performed over months to group similar months, then K_2 clusters are obtained from the previous K_1 datasets to detect daily patterns and finally, with the K_2 datasets, similar hours are grouped independently to describe hourly consumption under the schema profile K_1 (season K_1) during a day-type K_2. At this point, TLP is used to present the behavior of the hourly consumption to the final user. Even the TLP is a more complex tool of postprocessing, in this research, a basic application of TLP is conceptualized here; only using it to visualize the semantic of variables through the meaning of the colors. By this way, blocks of hours with high consumption are colored by red; medium consumption is identified by yellow color; and low consumption is presented with green color, respectively. Figure 1 summarizes the clustering process of the research.

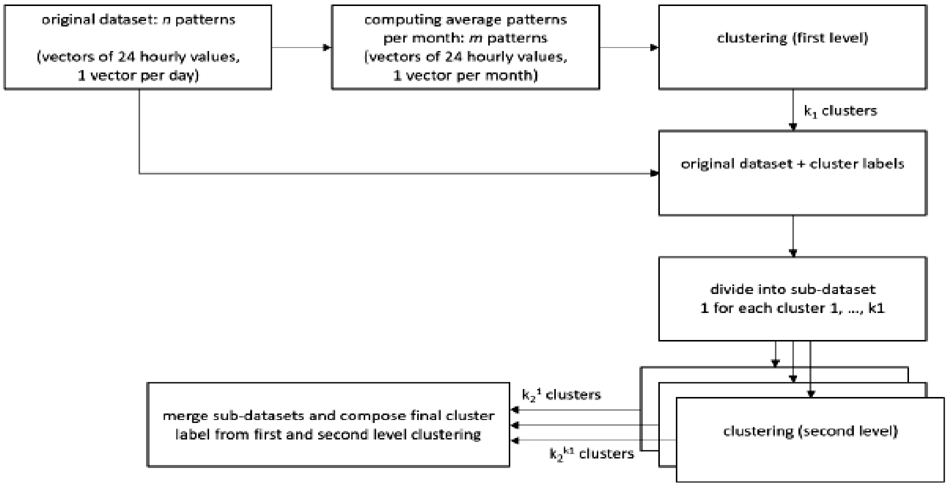
original dataset: *n* patterns
(vectors of 24 hourly values, 1 vector per day)
computing average patterns per month: *m* patterns
(vectors of 24 hourly values, 1 vector per month)
clustering (first level)
k_1 clusters
original dataset + cluster labels
divide into sub-dataset
1 for each cluster 1, …, k1
clustering (second level)
k_2^1 clusters
k_2^{k1} clusters
merge sub-datasets and compose final cluster label from first and second level clustering

Figure 1. Methodological approach: First level refers Seasons/months and second level is associated to typology of days.

3 APPLICATION DOMAIN

The research was carried out with a case study from Northern Mexico, specifically the region that includes the states of Sonora, Sinaloa and Baja California. Data from energy consumption was collected as time-series from January, 1st 2004 until August, 31st 2017. Each record (row) in the database contains the daily consumption of energy in kWh (kiloWatt hour) labeled as X1, X2, X3, …, X24, since X_i represents the consumption in kW of the hour i of the date in row j. At the same time, database contains the weather conditions for each day: maximum and minimum temperature in Celsius degrees, average of relative humidity in %, average of precipitation rate in mm and a categorical data which indicate special dates on the calendar related to Northern Mexico. The database was obtained from public data repositories of the National Center of Energy Control in Mexico.

4 EXPERIMENTAL RESULTS

Following the proposed methodology, data was pre-processed, and an exploratory analysis was also performed as a previous step. The software platform used is R (R Core team, 2017) connected with Power BI from Microsoft suite. From this analysis, authors and experts on the field agreed that the whole set of data would be reduced by considering only the years 2015, 2016 and 2017. This conclusion was taken after observing that the shape of the load curves remains the same all over the years from 2004 to 2017 and only changes in the measurement (increasing kWh) is detected. This insight was obtained from a previous analysis by using DTW (Dynamic Time Warping) and cosine similarity distances amongst the time-series database. Experts on the field explained that this is a direct consequent of the increasing in population of the studied geographical zone. Figure 2 shows the average daily consumption from 2004 to 2017 and the average hourly consumption from 2004 to 2017.

Clustering process with data from 2015 to 2017 was performed by applying Ward's method and cosine distance. The number of classes K was determined by optimizing the Calinski-Harabasz index by evaluating the more feasible cuts of the dendrogram in each step of the clustering process detailed in section 2.

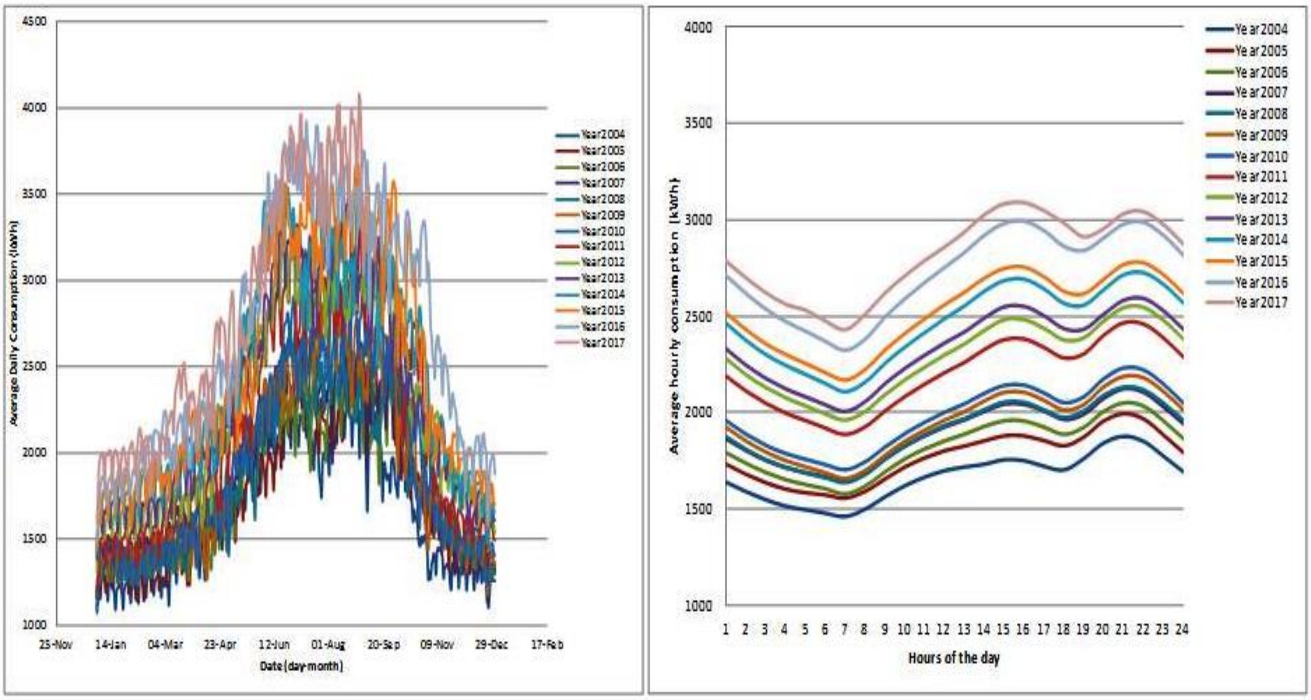

Figure 2. Average daily consumption (left) and average hourly consumption (right) from 2004 to 2017 in kWh.

Table 1 summarizes the results of the clustering process over the top/down levels in time-blocks.

Table 1. Summary of clustering results

Time-blocks Levels (top to down)	**Number of clusters obtained**	**Sub setting of the database**	**Interpretation of the clusters**
Months/Seasons	K1= 2	Original database was partitioned in two subsets for each value of K1. These datasets were the input for the next step.	Label 1 in K1 groups similar months in terms of energy consumption from June to October. Label 2 in K1 groups months from November to May. Labels of K1 indicate months/seasonal types.
Days	For the subset 1, the number of clusters was K2= 3 and for the subset 2, K2= 2, respectively.	Thus, subset 1 from the previous step was divided into 3 clusters and subset 2 into 2 clusters. Labels of K2 indicated typology of days.	Labels K1=1 and K2=1 indicates Sundays/holidays from June to October (Pattern 1), Labels K1=1 and K2=2 indicates Saturdays from June to October (Pattern 2) and finally, K1=1 and K2=3 groups Working days (Mondays to Fridays) during months from June to October (Pattern 3). Labels K1=2 and K2=1 indicates Sundays/holidays from November to May (Pattern 4) and Labels K1=2 and K2=2 indicates days from Mondays to Saturdays during the period November to May (Pattern 5).

Hours	K1=1 and K2=1, K3=4
	K1=1 and K2=2, K3=4
	K1=1 and K2=3, K3=4
	K1=2 and K2=1, K3=3
	K1=2 and K2=2, K3=4
		With the previous five load profiles (season/days) a final clustering process was executed to detect similar hourly blocks in each of the five load patterns resulting from clustering. K3 indicates the number of clusters associated to similar hourly blocks.	

After being labeled the clusters in terms of seasons and typologies of days, a brief description of them was obtained together with the experts to use this information for a better understanding of the behavior of energy consumption. The characterization of the energy consumption all over the year can be then summarized in five profiles or patterns. Pattern 1: energy consumption during “summer” (June to October) and Sundays or holidays. Pattern 2: energy consumption during summer and Saturdays. Pattern 3: energy consumption during working days and summer season. Pattern 4: energy consumption of Sundays or holidays during the rest of the year and Pattern 5 represents energy consumption during the rest of the days in the week (Mondays to Saturdays) during the rest of the year, respectively.

Finally, the hourly grouping for each of the five patterns is detailed in table 2.

Table 2. Hourly distribution for the five representative patterns of the case study.

Pattern	Groups of similar hours for each Pattern (season/type of day)
Pattern 1	[01:00-08.00],[09:00-16:00],[17:00-19:00],[20:00-24:00]
Pattern 2	[01:00-07.00],[08:00-16:00],[17:00-19:00],[20:00-24:00]
Pattern 3	[01:00-07.00],[08:00-16:00],[17:00-19:00],[20:00-24:00]
Pattern 4	[01:00-08.00] and [24:00], [09:00-19:00],[20:00-23:00]
Pattern 5	[01:00-07.00] and [24:00], [08:00-12:00],[13:00-18:00],[19:00-23:00]

5 DISCUSSION

Accuracy of clusters, calculated as the matching between labels of the clusters and the official calendar dates demonstrated the efficient classification (grouping) of the process. For the first (top) level, the matching related to month versus the two main seasons labeled by the experts was around 98%. The mismatch of 2% was due to some latter days of October (that could be interpreted as transitional days) and some latter days of May that could be grouped in some cases into the “summer” season. In relation with the typology of days, the match was around 92%, i.e., days matched with the actual day reflected on the date. The mismatching of 8% approximately was due to “special days” when temperature or another weather condition was out of the normal trends for the season. This group was extracted from data and labeled as “outliers” and they will be used as a “special class” in future works. Note that this “external data” helped to detect anomalous events and situations where the energy consumption is totally out of the thresholds no matter the type of days or season.

Figure 3 shows the characterization of the five main patterns. Note that these five profiles are clearly differentiated and provides the different behaviors according to season and type of days. It seems that this characterization is a remarkable insight that could be used as an input to forecast the current demand of energy by customizing the prediction according to the type of patterns. For instance, given a certain date, an easy routine could extract the type of day and its position related to the season (summer or rest of the year); this “labeled” input would activate the appropriate forecasting algorithm designed for that type of date according to the belonging to one of the five patterns obtained during this research. For this case study in Northern Mexico, summer season which groups months from June to October, can be labeled as the most challenging period in terms of production and load distribution due to the highest consumption occur during these months. Our characterization also shows that the shape of the patterns is similar (patterns 1, 2 and 3), but Sundays presents a “delay” in relation with the rest of the days at hour 7am to 8am. This happens probably because on Sundays, people are used to wake up later than the common working days. In addition, the energy consumption is lower during Saturdays and Sundays because many major consumers do not operate during the week-end (government, schools, industries and so on). On the other hand, the rest of the year period has clear lower energy consumption, this happens due to weather conditions are more comfortable from November to May and the phenomenon of “air-conditioning consumption” reduces drastically. The Shape of patterns 4 and 5 are smoother if compared with patterns related to summer period.

Hourly consumption similarity is a complementary and remarkable insight that supports operators to plan and scheduling load distribution in the power networks. Differentiated hourly blocks can show peaks/rush consumption hours that can be used to reduce or prevent overloading of the system. As shown in Table 2, the five patterns have differentiated distribution of hourly blocks in terms of their shapes, changes in time and measurements (kWh). For instance, in Pattern1 (Sundays in summer) the energy consumption of the 24 hours of this type of pattern can be grouped in four blocks (X1 to X8; X9 to X16; X17 to X19; X20 to X24), i.e., hours from X1 to X8 have the same behavior amongst them in terms of energy consumption and different if compared with other of the three remaining groups calculated for this pattern. To summarize results at this level (postprocessing step); the basic TLP tool was built for hourly blocks. Two TLP´s were obtained aimed to establish separated energy consumption between seasons (from figure 3 is noted that exits a notorious difference in shape and measurement in kWh between the group of patterns 1, 2, 3 and the group of patterns 3, 4), one for patterns 1, 2 and 3 (due to their belonging to summer season) and a second one for patterns 3 and 4 (rest of the year). Figure 4 shows the TLP’s associated to hourly blocks into the five profiles that describe the energy consumption. Selection of colors for each group was done with the help of the experts and conditional statistics. Here, green color represents hours with low consumption, yellow is medium and red color cells can be considered as peak hours or periods with high consumption trends.

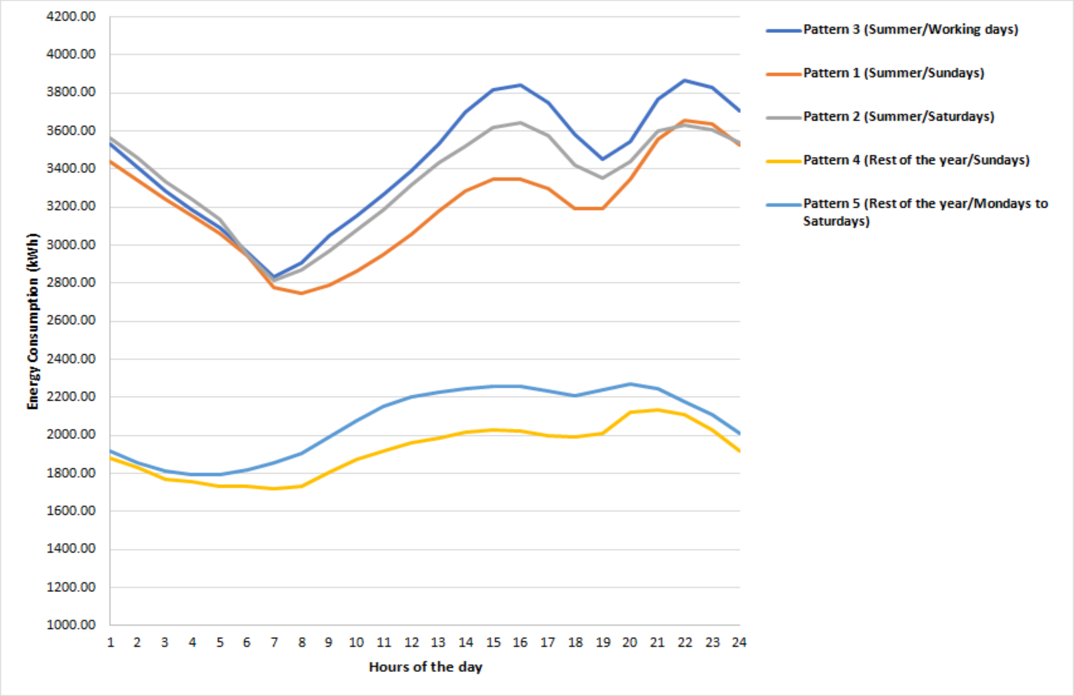

Figure 3. Representative curves of the five load profiles: Average Hourly Consumption in kWh.

Hours Blocks/Patterns	X1 to X7	X8	X9 to X16	X17 to X19	X20 to X24
Pattern 1	🟩	🟩	🟢	🟨	🟥
Pattern 2	🟩	🟨	🟨	🟧	🟥
Pattern 3	🟩	🟨	🟨	🟧	🟥

	X1 to X7	X8	X9 to X12	X13 to X18	X19	X20 to X23	X24
Pattern 4	🟩	🟩	🟨	🟨	🟨	🟥	🟩
Pattern 5	🟩	🟨	🟨	🟧	🟧	🟥	🟩

Figure 4. TLP´s: Hourly block groups for each of the five profiles.

TLP’s of figure 4 groups the similar hourly-blocks in a friendly context to facilitate the understanding of the profiles. Pure colors are associated to classes which clearly differences from the others. Degrade colors are associated to classes that could be interpreted as intermediate classes between two colors of the three basic colors in the TLP tool (pure green, yellow and pure red). Thereby, clearer green is an intermediate label from pure green to yellow and clearer red (almost orange color) is a label from yellow to pure red color, respectively. Under this color association TLP’s in figure 4 shows that patterns 2 and 3 have similar behavior in hourly blocks while pattern 1 has a different evolution of the energy consumption if compared with patterns 2 and 3 which belong to the same season (summer). In like manner, patterns 4 and 5 have quite different hourly evolution in energy consumption when comparing each other.

6 CONCLUSIONS & FUTURE WORKS

Results were discussed with experts on the field. The characterization, qualitative and quantitative, of the five main patterns changed the way as experts were monitoring the behavior of the consumption during the year. Previously, they assumed two main seasons, as it happens in this research with K1, but in terms of typology of days, they managed six different types of days equally labeled in the two seasons (Mondays, Tuesdays to Thursdays, Fridays, Saturdays, Sundays and Holidays). Our clustering demonstrated that the current characterization of the final-users/experts could be improved with the new characterization obtained in this research. It was tested that the top/down approach dealt with a better understanding of the load curves during the year. In addition, hourly blocks improved the knowledge of the energy consumption acquired so far. Nevertheless, as shown in figure 4, TLPs associated for each pattern in relation of the hourly consumption presented pure and non-pure colors, so they are intermediate TLPs that need to be improved. For this reason, a first future line is aimed to re-build new TLPs by considering the grade of uncertainty within the profiles. These new TLPs will be called annotated TLPs (a-TLPs) as suggested in Gibert and Conti, 2015. The a-TLP is an enriched tool to manage the intrinsic uncertainty related with prototypes. The aTLP handles uncertainty through a quantification of the prototypes' purity based on the coefficient of variation (CV) and an associated color-based uncertainty model, with two dimensions – tone and saturation – representing nominal trend and purity of the prototype.

A second future line, given the characterization obtained in this research, refers to the developing of a forecasting engine to predict energy consumption. The process will use the first six current hours at the beginning of each day (which commonly are "stable" hours due to their association with night/early morning period) as inputs to forecast the next 18 hours of the day. The idea deals to "customize" forecasting models (classical versus machine learning algorithms) for each pattern obtained so far. This approach has been applied in previous studies developed by the authors in power systems and urban water distribution networks with good results.

REFERENCES

Candelieri, A., 2017. Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection. Water. 9 (3), 224-243.

Candelieri, A., Archetti, F., 2014. Identifying Typical Urban Water Demand Patterns for a Reliable Short-Term Forecasting – The ICeWater Project Approach. Procedia Engineering. 89, 1004-1012.

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics.3 (1), 1-27.

Deshani, K., Attygalle, M, Hansen, L., 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.

Fayyad, U., 1996. From Data Mining to Knowledge Discovery: An overview. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

Gibert K., Garcia-Rudolph A., Garcia-Molina, A., Roig-Rovira, T., Bernabeu, M., Tormos, J.M., 2008. Response to TBI-neurorehabilitation through an AI & Stats hybrid KDD methodology. Medical Archives. 62(3), 132–135.

Gibert, K., Conti, D., 2015. aTLP: A color-based model of uncertainty to evaluate the risk of decisions based on prototypes. AI Communications. 28, 113-126.

Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and techniques, 3rd Edition, Morgan Kauffman.

Hernandez, L., Baladron, C., Aguiar, J.M., Carro, B., Sanchez, A., 2012. Classification and Clustering of Electricity Demand Patterns in Industrial Parks. Energies. 5, 5215-5228.

Martinez-Alvarez, F., Troncoso, A., Riquelme, J.C., Aguilar-Ruiz, J.S., Petermann, C., 2011. Energy Time Series Forecasting Based on Pattern Sequence Similarity. IEEE Transactions on Knowledge and Data Engineering. 23(8), 1230–1243.

Plaza, M.A., Conejo, A.J., Prieto, F.J., 2005. Multimarket Optimal Bidding for a Power Producer. IEEE Transactions on Power Systems. 20(4), 2041–2050.

R Core Team, 2017. R: A language and environment for statistical computing. Vienna, Austria.

Rhodes, J.D., Cole, W.J., Upshaw, C.R., Edgar, T.F., Webber, M.E., 2014. Clustering Analysis of Residential Electricity Demand Profiles. Applied Energy. 135, 461–471.

Servidone, G., Conti, D., 2016. Discovering and labeling of temporal granularity patterns in electric power demand with a Brazilian case study. Pesquisa Operacional. 36(3), 575-595.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Statis. Ass. 58, 236-244.