ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 7 N. 2 (2018), 53-69
eISSN: 2255-2863
DOI: http://dx.doi.org/10.14201/ADCAIJ2018725369

Customized normalization clustering methodology for consumers with heterogeneous characteristics

Catarina Ribeiro[a, b], Tiago Pinto[a], Zita Vale[c] and José Baptista [b, d]

[a] Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development (GECAD), Institute of Engineering, Polytechnic of Porto (ISEP/IPP), Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal
[b] UTAD – University of Trás-os-Montes e Alto-Douro, Vila Real, Portugal
[c] Polytechnic of Porto, Porto, Portugal
[d] CPES - INESCTEC
{acrib, tmcfp, zav}@isep.ipp.pt; baptista@utad.pt

KEYWORD

Clustering; Data Mining; Smart Grid; Consumption Profile, Dynamic Tariffs, Prosumers.

ABSTRACT

The increasing use and development of renewable energy sources and distributed generation, brought several changes to the power system operation. Electricity markets worldwide are complex and dynamic environments with very particular characteristics, resulting from their restructuring and evolution into regional and continental scales, along with the constant changes brought by the increasing necessity for an adequate integration of renewable energy sources. With the eminent implementation of micro grids and smart grids, new business models able to cope with the new opportunities are being developed. Virtual Power Players are a new type of player, which allows aggregating a diversity of entities, e.g. generation, storage, electric vehicles, and consumers, to facilitate their participation in the electricity markets and to provide a set of new services promoting generation and consumption efficiency, while improving players` benefits. This paper proposes a clustering methodology regarding the remuneration and tariff of VPP. It proposes a model to implement fair and strategic remuneration and tariff methodologies, using a clustering algorithm, applied to load values, submitted to different types of normalization process, which creates sub-groups of data according to their correlations. The clustering process is evaluated so that the number of data sub-groups that brings the most added value for the decision making process is found, according to the players characteristics. The proposed clustering methodology has been tested in a real distribution network with 30 bus, including residential and commercial consumers, photovoltaic generation and storage units.

1. Introduction

The power sector has been completely revolutionized by the emergence of liberalized electricity markets, aiming to improve the system's efficiency while offering economic solutions. Governments all over the world are trying to increase the use of renewable energy over the last few decades (Sharma K.C. *et al*, 2014). The sector's restructuring process brought out several challenges, requiring the transformation of the conceptual models that previously dominated the power sector (Sioshansi, F.P., 2013). The privatization, liberalization and international integration of previously nationally owned systems are some examples of the transformations that have been applied, along with the energy markets evolution into regional and continental scales, supporting transactions of huge amounts of electrical energy and enabling the efficient use of renewable based generation in places where it exceeds the local needs (Sharma K.C. *et al*, 2014), (Sioshansi, F.P., 2013).

Despite the increase in renewable energy sources production, which is a favourable scenario for the development of distributed generation (DG), important difficulties arise with their large scale integration. The coordination between technical and economic issues is much more complex in the present context because the instability of technologies like wind or solar plants. Electricity markets operation has to consider the physical constrains of power systems, market operation rules and financial issues such as the resources dispatch, the participation of small producers in the market and high costs of maintenance, all problems that must be solved in order to make DG a real advantage (M.Shadidehpour, *et al*, 2002).

Potential benefits will depend on the efficient operation in the market and, on the other hand, in the remuneration of aggregated players. Important developments concerning electricity market players modelling and simulation including decision-support capabilities can be widely found in the literature (I.Praça, el al, 2003).

Much like electricity markets, subsystems of the main network are rapidly evolving into a reality, coordinating these entities is a huge challenge that requires the implementation of distributed intelligence, potentiating the concept of Smart Grid (SG) (M.Shadidehpour, *et al*, 2002), (Blumsack S. and Fernandez A., 2012). However, the two concepts are not converging towards common goals and technical and economic relationships are addressed in an over simplistic way. Present operation methods and electricity markets models do not take full advantage of installed DG, yielding to inefficient resource management that should be overcome by adequate optimization methods (Sousa T., *et al*, 2012). Player aggregating strategies allows players gaining technical and commercial advantages, individuals can achieve higher profits due to specific advantages of a mix of technologies to overcome disadvantages of some technologies. The aggregation of players gives rise to the concept of Virtual Power Player (VPP) (Z. Vale *et al*, 2011). VPP are heterogeneous entities considered to enable widespread inclusion of distributed energy resources (DER), energy storage systems (ESS), electrical vehicles with gridable capability (V2G) and consumers, considering demand response (DR) programs. VPPs' participation in electricity markets provides a set of new services to promoting efficiency in generation and consumption and improving players' benefits (T.Pinto *et al*, 2009). Each aggregated player has its individual goals; hence the VPP should conciliate all players in a common strategy, able to allow each player to pursuit its own objectives (Z. Vale e*t al*, 2010).

This article presents a data mining methodology, based on the application of a clustering process to load values, submitted to different types of normalization process, which groups the typical load profile of the consumers of a SG according to their similarity. The separation of consumers in different groups allows proposing specific consumption tariffs to each group, so that consumers' load profile is taken into account to meet the objectives of the SG aggregator. The proposed clustering methodology is tested in a real distribution network with 30 bus, including residential and commercial consumers, photovoltaic generation and storage units. This work thus proposes a customized normalization method to treat data before it is used by the clustering process. After this introductory section, Section 2 introduces the concept of SG and electricity markets simulation, section 3 gives details of Remuneration and tariff mechanism -RemT, the tool that is being development for decision support of electricity markets remuneration and tariff. In section 4 are presented the data normalization methods and the clustering process k-means. A case study is presented in section 5 based on real consumption data collected from a real SG with 82 consumers, the clustering process is applied after the treatment of the data through several methods. Final conclusions are featured in section 6.

2. Smart Grid and Electricity Markets simulation

Many works have been developed using simulators to model the complex interaction between electricity market players. Successful examples sustain the fact that a multi-agent system (MAS) with adequate simulation abilities, is the best approach for simulating electricity markets. The Multi-Agent Simulator for Competitive Markets – MASCEM, (T.Pinto *et al*, 2011) is a platform that simulates several electricity market types, while providing decision support to players' actions. This type of simulators are able to represent market mechanisms and players' interactions. However, for them to be valuable decision support tools in foreseeing market behaviour, they need to be used in testing adequate and realistic scenarios. Real data analysis by means of a knowledge discovery process will be a crucial step forward to assure that MASCEM agents exhibit adequate profiles and strategies.

Multi-Agent Smart Grid simulation Platform – MASGriP (Oliveira P. *et al*, 2012), simulates, manages and controls the most important players acting in a Smart Grid environment. This system includes simulated players, which interact with agents that control real hardware. The considered players include operators, and energy resources, such as several types of consumers, producers, electric vehicles, among other. Aggregators are also considered, namely: VPPs and Curtailment Service Providers (CSP) (C. Kieny *et al.*, 2009). These players introduce a higher level of complexity to the management of the system. Joint simulations of MASCEM and MASGriP enable a simulation environment that includes the participation of Smart Grid players in electricity markets, or even internal Smart Grid markets, using complex markets models provided by MASCEM.

The decision support mechanism RemT that is being developed to support the VPP actions in the scope of MASCEM, to define the best tariff and remuneration to apply to each of the aggregated players, regarding the VPP objectives and the individual goal of each aggregated player.

3. Decision support tool for electricity markets remuneration and tariff definition RemT

Remuneration and Tariff Mechanism – RemT, (C. Ribeiro *et al.*, 2013)(C. Ribeiro *et al.*, 2015) is a decision support mechanism that is being developed to support the VPP actions in the definition of the best tariff and remuneration to apply to each of the aggregated players, regarding the VPP objectives and the individual goal of each aggregated player. VPPs in the scope of MASCEM use RemT to remunerate aggregated players, according to the results obtained in the electricity market, the penalties for breach of contract, contracts established to guarantee reserve, demand response programs and incomes of aggregated consumers. The definition of remuneration and tariffs is based on the identification of players' types and on the development of contract models for each player type. This considers players with a diversity of resources and requirements, playing several distinct roles (a player can be a consumer, a producer and can be responsible for one or several electrical vehicles). The terms for new contracts and the best strategies for each context are determined by means of machine learning methods and data-mining algorithms (Z. Vale *et al.*, 2011). The definition process is presented in Figure 1.

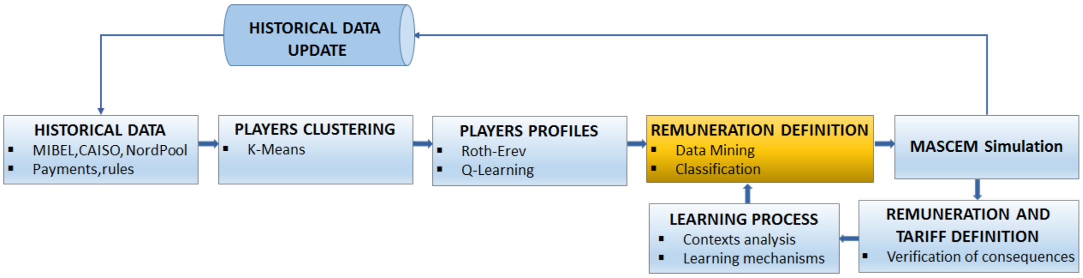

Figure 1: RemT definition process

The establishment of remuneration and tariffs is based on the identification of players' types and on the development of contract models for each player type. The remuneration mechanism is executed after the closing of the markets and of the bilateral negotiation process. The results of remuneration and tariff process are used in a learning algorithm to improve the quality of the remuneration process, with implications to the market bidding process.

3.1. Historical Data Module

The historical data base is updated after each simulation, including the agent's results in the market. This historic log is used by the VPP to get information regarding the agents' behaviour. With this information VPPs are able to take conclusion about several very important aspects, such as understanding if an agent usually meets the contracts agreements it committed to; or if agents have a production type that the VPP finds interesting to aggregate, etc.

The data consists in information about real transactions of consumers, energy producers, VPPs, real payments, contracts and rules define for each type of market. The energy markets used as sources are (MIBEL, 2017), (EPEXSPOT, 2017), (Nord Pool Spot, 2017), (CAISO, 2017).

3.2. Players Clustering Module

Players clustering is very important, in order to optimize the remuneration and tariff strategies. A classification method is required to obtain a rule set that can be used to classify the aggregated players in the defined clusters. In order to identify interesting relations in players' data base, a clustering mechanism is used to group the Players, in different groups, according to their characteristics. The clustering is performed using the K-Means algorithm provided by MATLAB. The objective function (1):

$$J = \sum_{j-1}^{k} \sum_{i-1}^{n} ||x_i^{(j)} - c_j||^2 \tag{1}$$

3.3. Players Profiles Module

The methodology takes into account relevant factors concerning each player, their contribution to the aggregator, and the market and operation context. These factors include the VPP market results, contract portfolio and clauses, players' behaviour and classification, Locational Marginal Prices (LMPs), participation in ancillary services and demand response programs, players' consumption profile, V2G usage profile and requirements, and parking and stations remuneration and tariffs.

The construction of profiles is made using different research techniques in the historical data of the agents, some learning mechanisms are also used.

Roth-Erev Reinforcement Learning Algorithm, the main reasoning to use this technique is that the tendency to perform an action should be strengthened, or reinforced, if it produces favourable results and weakened if it produces unfavourable results, the main principle is that, not only are choices that were successful in the past more likely to be employed in the future, but similar choices will be employed more often as well, this is referred to as law of effect (Erev I. and Roth A., 1998). This algorithm uses an experimentation or regency parameter, which defines the weight that the past experience will have on a subjects learning. The algorithm is able to successfully track the observed intermediate-term behaviour of human subjects over a wide variety of multiagent repeated games with unique equilibrium achievable, using stage-game strategies.

This algorithm allows the autonomous establishment of an interactive action policy, converges to the optimal proceeding when the learning state-action pairs Q is represented in a table containing the full information of each pair value (T. Pinto *et al.*, 2011). The algorithm is able to learn a function of optimal evaluation over the whole space of state-action pairs *s x a*. The Q function performs the mapping in the way represented in (2):

$$Q: s \times aU \tag{2}$$

Where *U* is the expected utility value when executing an action *a* in the state *s*. As long as the action states do not omit relevant information, nor introduce new information, once the optimal function *Q* is learned, the agent will know precisely which action results on the higher future reward, in a particular situation *s*. The *Q(s, a)* function, regarding the future expected reward when action *a* is chosen in the state *s*, is learned through try and error, following the equation (3):

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + a[r_t + \gamma V_t\,(s_{t+1}) - Q_t(s_t, a_t)] \tag{3}$$

Where *a* is the learning rate; *r* is the reward or cost resulting from performing the action *a* in the state *s*; *y* is the discount factor; and *U* is the state *s* utility resulting from action *a*, represented in function (4), obtained using the *Q* function learned so far.

$$U_t(s_{t+1}) = max_a\, Q(s_{t+1}, a) \tag{4}$$

The *Bayes Theorem* is another learning algorithm used because, one of this theorem advantages is its applicability to be used as a reinforcement learning algorithm. This applicability is based on the use of a probability estimation to determine which of the different alternatives presents a higher probability of success in each context, therefore being considered as the most appropriate approach. Bayesian networks have been developed to facilitate the task of prediction and abduction in artificial intelligence systems. Simplistically, these networks, also known as *causal networks*, or *probabilistic networks*, are graphic models for uncertainty based reasoning (Dore A. and Regazzoni C., 2010).

3.4. Remuneration Definition Module

After the historical data processing and players profiles definition phases, the remuneration and tariff definition process considers a data mining (A. Chrysopoulos AC, *et al.*, 2009) based task.

The extraction of knowledge from these data is based on the process of knowledge discovery in databases. The data mining process includes the use of algorithms to discover patterns among the data, following a similarity criterion. This has the purpose of assembling the data in such a way that each cluster contains objects with a high similarity among them and a high imparity with objects of other clusters. After the clustering step, the classification model is implemented with the main goal of creating a set of classification rules. These are defined according to a given scenario, taking into account the set of input attributes. This task enables the VPP to analyse and understand the profiles of its aggregated players so that the strategies can be adapted to each player. A clustering methodology is also used for this purpose.

The remuneration and tariff strategies definition and modelling is made according to the profiles established in the clustering process, taking into account the market rules, system characteristics, and real time system operation. The VPP, using this methodology, defines strategies for producers' remuneration, storage units' remuneration, consumption tariffs and V2G remuneration and tariffs.

These strategies consider the benefits both for the players, and also concerning the VPP profits. On one hand, the main goal of the VPP is to maximize its profits, which requires selling at high prices, buying at low prices and trading large quantities of energy. However, in a competitive environment, aggregated players can eventually act directly in the electricity market or aggregate themselves to another VPP. In other words, VPPs strategies must take into account each player satisfaction, considering not only its own benefits but also the benefits for their aggregated players.

The use of heuristic techniques is important to simultaneously optimize the VPP's and the players' profits (Erev I. and Roth A., 1998). The problem is a contradictory multi objective problem, in which the solution depends on the strategies used by the VPP. A machine learning based method (Z. Vale *et al.*, 2011) determines the best remuneration and tariff strategy.

3.5. Remuneration and Tariff Evaluation Module

The VPP chooses the contract that best suits the agent, considering the terms of the contract according to the importance of the agent. If the agent is considered to be important, the VPP can offer contracts with better terms in order to attract the agent to its aggregation. On the other hand, if the agent is not that important, the remuneration and the termination clauses of the contract will be lower. This way, the interests of either the VPP or aggregated agents are safeguarded in a more transparent and fair way. MASCEM is used to simulate the use of the obtained contracts and tariffs.

The implementation of the RemT module in MASCEM allows the simulation and validation of the obtained tariffs and remuneration contracts and when they are defined, a large set of case studies can be simulated, in order to obtain the required data for the machine learning module.

This module allows the comparisons of actions suggested by the mechanism, with those there are applied in reality, enabling the verification of consequences. Also allows the analysis of the results from applying the various alternatives, both from the standpoint of market and players. This analysis can be global, for all players, using a formula that combines the results of all players, and returns a global value of utility. In other hand, the analysis can be made for each type of player, resulting for the clustering mechanism.

3.6. Learning Process

Market simulation considers the players remuneration, where the inclusion of the remuneration and tariff process in MASCEM allows the validation of the adopted strategies for each VPP in a competitive environment. The remuneration process cannot be treated separately from market results and from operation results. For example, if a VPP does not sell any of its available power, the VPP does not pay anything to players, or simply base amount, according to the established contracts. On the other hand if a VPP sells a quantity of energy and the producers do not supply the amount of power they have committed, the VPP needs to adapt the remuneration strategy to this situation, while penalizing such producers. The VPP needs to consider all these situations in simulation scenarios and adapt remuneration strategies according to the established contracts and the market rules.

The learning module considers context analysis, using the historical behaviour of each player in different situations, such as different days of the week, different months, different hours in each different days of the week, special days with special events, and the different countries were players act or are located. This module also considers adapted learning mechanisms depending on the different referred contexts.

4. Data normalization methods and clustering algorithm for decision support tool

For certain algorithms, the reduction of the data to the same scale is fundamental. This need arises from the fact that certain non-parametric algorithms assume implicitly that the distances in different directions of the input space have the same weight. Indeed, without recourse to normalization (or standardization as it is also known in the scientific community (Gan *et al*., 2007), variables with large numerical values can dominate the effects of variables with smaller but equally important values, in definition of the model, such as age versus salary or power value of a set of high or medium voltage consumers, with consumption values perfectly unequal in terms of value amplitude.

Thus, the choice of the normalization factor should be made considering the type of data available, the type of analysis to be performed, as well as the type of data mining algorithm used. Several different normalization factors have been described for obtaining typical consumer profile.

Gianfranco Chicco (Chicco, *et al*., 2003) uses models based on cluster analysis. A consumer characterization model is presented for the study of tariff options. The characterization of consumers is based on different classes represented by their typical daily consumption profile. Consumption profiles are performed through a clustering algorithm. In this work, the authors propose the use of two measurement indices, the Mean Index

Adequacy (MIA) and the Clustering Dispersion Indicator (CDI), which help to decide the optimal number of groups, as well as assess the partition quality generated by the algorithms clustering.

In (Ramos *et al.*, 2012) a comparison between 5 clustering algorithms is presented in order to identify typical consumption profiles from a real database of medium voltage consumers. The partition performance of 5 different algorithms was compared and evaluated through the use of 12 measurement indices, namely Hubert Statistic, Normalized Hubert Statistic, Dunn index, Davies-Bouldin index, Squared Error index, Xen-Beni cluster validity index, Root-mean-square standard error, R-squared index, SD validity index, and finally Point Symmetry index.

Some indexes have criteria of maximization and others of minimization. Considering the validation of the indices, the algorithm that obtained the best results of grouping was the K-means, thus, the results of the characterization of these consumers point to the use of the algorithm K-means.

4.1. Clustering approach

The data mining methodologies presented in this paper is based on the application of a clustering process, which groups the typical load profile of the consumers of a SG according to their similarity. A wide variety of clustering algorithms can be found in the literature and unfortunately, there is no single algorithm that can by itself, discover all sorts of cluster shapes and structure (Anil K. Jain *et al.*, 1999).

K-means (Anil K. Jain, 2010), has been used, as it proves to be a robust model for distinct applications: K-means minimizes the distance from each point to the centre of the respective cluster, as defined in (5).

$$\min \sum_{i=1}^{k} \sum_{x \in c_i} \left\| x - \mathrm{m}_i \right\|^2 \tag{5}$$

Where μ_i is the mean of points in C_i, *i.e.* the cluster *centroid.* To determine the quality of the division of players into different clusters the clusters validity indices MIA and CDI (Chicco *et al.*, 2009) have been used, as formalized in (6) and (7) respectively.

$$MIA = \sqrt{\frac{1}{K} \sum_{k=1}^{K} d^2\left(x^{(k)}, \mathrm{m}^{(k)}\right)} \tag{6}$$

$$CDI = \frac{\sqrt{\frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{2.n^{(k)}} \sum_{n=1}^{n^{(k)}} d^2\left(x^{(m)}, \mathrm{m}^{(k)}\right) \right]}}{\sqrt{\frac{1}{2K} \sum_{k=1}^{K} d^2\left(x^{(k)}, R\right)}} \tag{7}$$

Where *d* represents the Euclidian distance between two points, and *R* is the representative load profile of all consumers. This indices represent distances, the smaller (or greater) is the MIA and CDI value, it indicates more (or less) compact clusters. To facilitate the analysis of results, we will consider that the higher the value of the MIA and CDI, the larger the error associated with this cluster.

4.2. Normalization methods and data treatment

Given that typical consumption patterns are to be found in the consumers, the reduction of their representative load diagram to a single scale is fundamental. It will allow to make them comparable to each other, otherwise consumers with higher consumption values could dominate the effects of load diagrams belonging to customers with lower consumption value, but still with the same behaviour. The standardization factor must be carefully chosen taking into account the type of data available, the analysis that is intended to be carried out, as well as the type of final results desired to be obtained. Based on previous studies concerning the characterization of electric energy consumers (C.Ribeiro *et al*, 2016), the maximum power value of the representative load diagram of each consumer was selected as a normalization factor. With the application of the normalization factor, all load diagrams assume the same order of magnitude, belonging to the interval [0,1], being able to be used by clustering algorithms, in order to be grouped according to a criterion of similarity, without losing information related to differences between amounts of consumption among consumers.

Analysing the results of previous works (C. Ribeiro *et al.*, 2013), (C.Ribeiro *et al*, 2016), is possible to verify that aggregation strategies have very good results and are very useful, because they provide a good separation according to what is intended.

The non-normalization grouping process has led to a clear separation between different consumers types, as it considers the absolute consumption amounts in the clustering process. The normalized data, used as formalized in (8) and (9), reveals a separation through consumption profiles, although it is not able to consider the differences in consumption quantity. *L* is the value of load.

$$N_{c,h}=\frac{L_{c,h}}{ML_c},\forall c\in co \tag{8}$$

$$ML_c=\max(L_c),\forall c\in co \tag{9}$$

Where *N* is the common normalized load, for each consumer *c*, for each hour *h, a*nd *co* is the set of all considered consumers. *ML* is the largest consumption value, of the consumer *c*, considering all hours.

To improve the results achieved in the previous works, the data were treated in other forms and were tested, the called average process, difference process and customized normalization process are introduced.

4.2.1 Average process

In this process average is made between regular data and normalized data, for each type of consumer. The power value to be considered in the clustering process was found by averaging between the regular load values and the normalized load values for each specific load in the 24 periods, it is formalized in (10), where *A* is load with average process data treatment.

$$A_{c,h}=\frac{L_{c,h}+N_{c,h}}{2},\forall c\in co \tag{10}$$

4.2.1. *Difference process*

In difference process the value of the micro production *P*, generated in the bus associated to a load, was subtracted from the value of the consumption of that load. This calculation is performed for loads from 1 to 17, loads which are on buses with associated micro production, for each specific load in the 24 periods. It is formalizes in (11) and (12).

$$T_{c,h}=\left|L_{c,h}-P_{c,h}\right|,\forall c\in co \tag{11}$$

Where *T* corresponds to the load were consumption was subtracted from micro production, for each consumer *c*, for each hour *h*. *SN* is the load with a different normalization process, for each consumer *c*, for each hour *h*. *SML* is the largest consumption value recorded for all consumers at the time *h*, it is formalized in (12).

$$SN_{c,h}=\frac{T_{c,h}}{SML_h},\forall c\in co \tag{12}$$

Two independent variables are subtracted to accentuate the difference between classes of loads. This is because, apparently, local production depends on the class of loads, see (Canizes B *et al.*, 2015), residential houses produce more throughout the year than they consume whereas it is the opposite for commercial building, and residential building produce as much as they consume. This supports the use of the proposed method.

4.2.1. *Customized normalization*

This method normalizes data using each consumer's load value at each period divided by the largest recorded value of all loads in all periods, it is formalized in (13) and (14).

$$SN_{c,h} = \frac{L_{c,h}}{SML_h}, \forall c \in co \tag{13}$$

$$SML_h = \max(L_{co,h}), \forall c \in co \tag{14}$$

Where *SN* is the load with a different normalization process, for each consumer *c*, for each hour *h*. *SML* is the largest consumption value recorded for all consumers at the time *h*.

The proposed customized normalization method aims to combine the advantages of both previous approaches (using non-normalized data, and regular normalization), so as to achieve consumer groups that capture both differences in the quantities of consumption, and also the trends of consumer profiles along the hours. The clustering process takes into account the tendency of the consumption values trough the time, regardless of its absolute amount. This separation is very important, according to different consumers' types and profiles, it works as a base for personalized and dynamic consumption tariff definition. Using this approach ensures that the data are also normalized in a range between 0 and 1, but without losing information related to differences between amounts of consumption among consumers. While using the regular normalization, the value 1 is attributed to the greater consumption value of each consumer (thus both consumers with large and small values will always have one value of 1 in a certain hour), using the customized normalization method, only the largest consumer of all, will have a value of 1. The smaller consumers will have normalized values with smaller values, proportional to the difference between the quantities of consumption of that consumer and the largest consumer in each hour. Thus normalization is still made between 0 and 1, but there is visible difference between higher and lower consumption among different consumers, and the evolution of consumption of each consumer profile is also captured

5. Case Study

This case study intends to show the adequacy of the proposed customized normalization clustering methodology to solve the problem of remuneration of players with heterogeneous characteristics and behaviours. In order to test the adequacy of the method, a clustering algorithm has been applied, concerning the consumption data of a total of 82 consumers (8 residential houses, 8 residential buildings with 72 loads, and 2 commercial buildings). Data has been collected from a real distribution network throughout one year. The Smart grid accommodates distributed generation (photovoltaic and wind based generation) and storage units, which are integrated in the consumption buildings. The accommodated photovoltaic generation, wind based generation and storage units are related to the building installed consumption power, according to the current legislation in Portugal. Further details on the considered distributed network can be seen in (Canizes B *et al.*, 2015). The distribution network of the SG is represented in Figure 2.

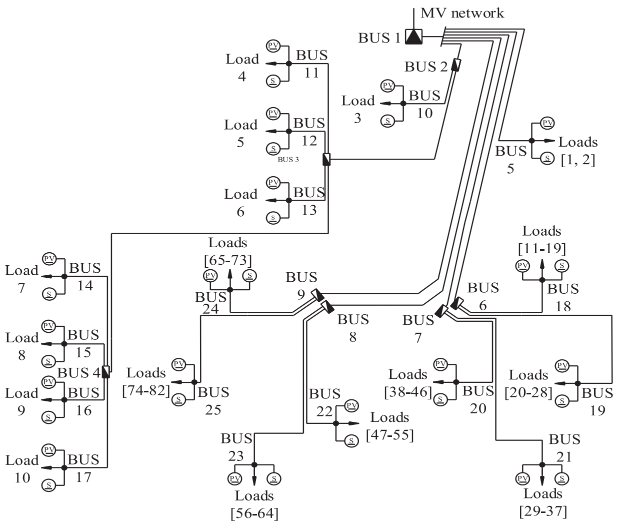

Figure 2: Distributed network for the SG simulation

The K-means algorithm has been used to perform the clustering process using non-normalized values of load (section 5.1), and also normalized values, using the regular normalization method (section 5.2), the proposed average process (section 5.3), difference process (section 5.4) and customized normalization method (section 5.5).

5.1. Non-normalized data

The clustering process is performed for different numbers of clusters, from 2 to 12, the maximum number of 12 clusters was defined in order to have a data set that was not too large to analyse and allow to reflect different consumption profiles. In order to enable grouping consumers according to the similarity of their consumption profiles, in order to support the definition of specific tariffs that are suited for each of the consumer groups. These number of clusters were defined not to create a broad portfolio of tariffs. From (C. Ribeiro *et al.*, 2015) it has been concluded that, by analysing MIA and CDI results from the clustering of non-normalized data, the best clustering results are achieved with the use of 3 clusters, as the clustering error is minimal. MIA and CDI results are presented in Figure 3.

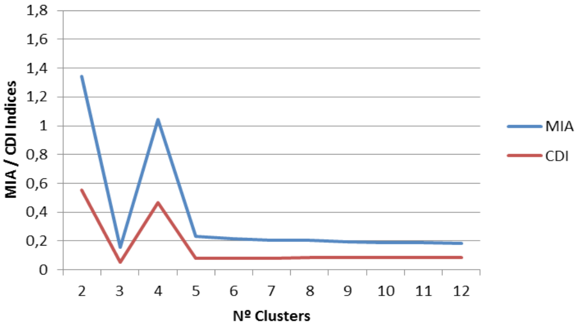

Figure 3: MIA and CDI results for non-normalized load, for 24 periods

When using 2 clusters, a clear separation of residential houses and buildings from commercial buildings is visible. It is also visible that the two commercial buildings (corresponding to loads 1 and 2) have been allocated to cluster 1, and the rest of the loads, corresponding to residential consumers, have been aggregated in cluster

2. This can be observed in Figure 4 which presents the load profiles of consumers that have been grouped in cluster 1 and in cluster 2 using the non-normalized data.

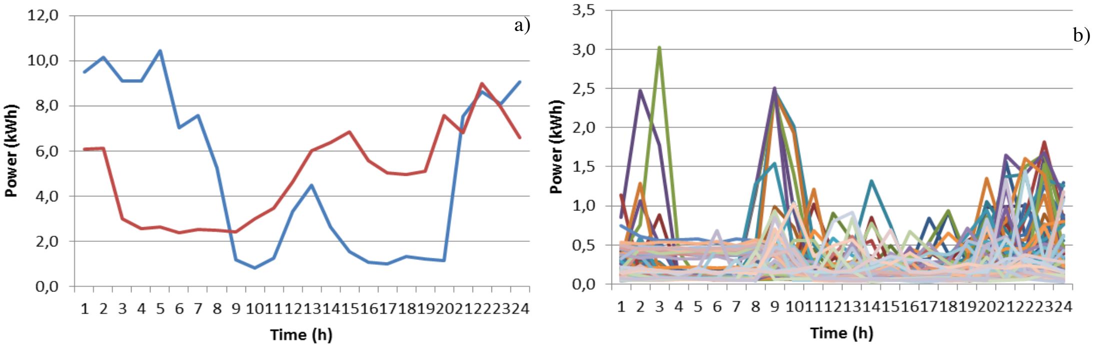

Figure 4: Consumption profile of loads allocated to: a) cluster 1; b) cluster 2

From Figure 4 it is visible that cluster 1 includes the two commercial buildings, with very distinct load profiles, and cluster 2 includes all the residential buildings and houses. When considering the grouping process with 3 clusters, the difference is that there is still a separation from residential houses and buildings to the commerce. However, in this case the two types of commercial buildings are also separated, as they present very different load profiles.

5.2. Normalized data

In the second clustering process, regular normalized data were used. The normalization was made considering each type of consumer. The value of load corresponding to each period was divided for the maximum value register in that specific load in the 24 periods. When using normalized values, a more accentuated descent of the clustering error values is visible. In Figure 5 is visible that the descent in the error value is however, stable from the start, which hardens the identification of the optimal number of clusters that should be used.

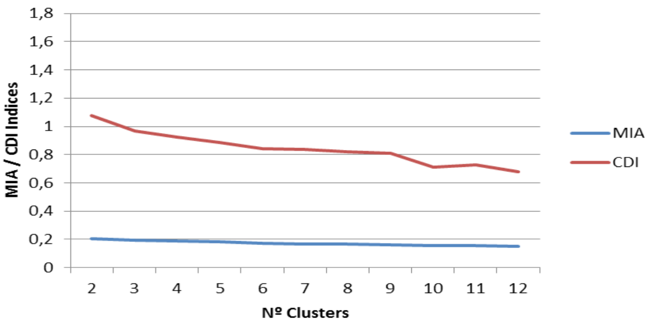

Figure 5: MIA and CDI results for normalized values for 24 hour period

For this reason it is not advantageous to use more than 2 or 3 clusters, since the use a larger number is not reflected by a significant gain in clustering error. By analysing the results of the clustering process with 2 clusters, it can be seen that the separation is not as clear as it was with non-normalized values. The two commercial buildings corresponding to load 1 and 2, were aggregated in different clusters, together with several residential consumers. However, the clustering process with normalized values has better results from the load profile separation stand point, as can be seen from Figure 6, which presents the allocation of the consumers to the different clusters, when considering normalized data and 2 clusters.

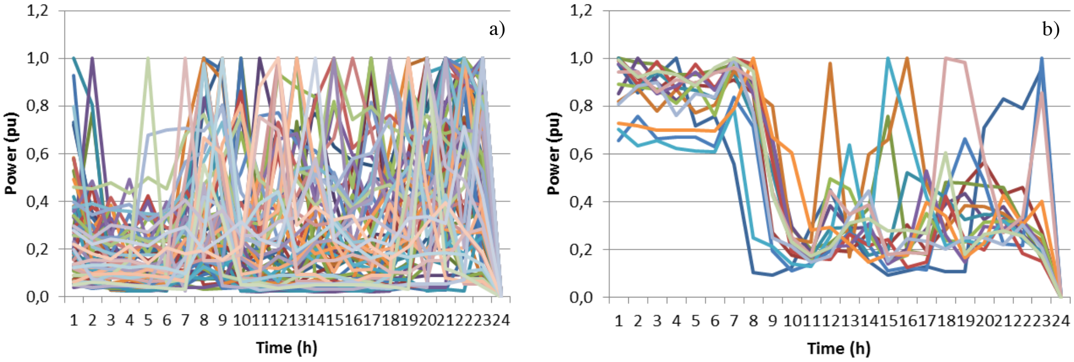

Figure 6: Consumption profile of loads represented in: a) cluster 1; b) cluster 2

From Figure 6 it is visible that although the consumer types cannot be separated correctly with this approach as occurs when using non-normalized data (Figure 3), the separation of the load profiles is more evident in this case, since profiles are grouped independently from the gross amount of consumption itself.

5.3. Average process

In the third clustering process was considered regular data and normalized data, for each type of consumer. The power value to be considered in the clustering process was found by averaging between the regular load values and the normalized load values for each specific load in the 24 periods.

By analysing MIA and CDI in Figure 7 results from the clustering of average normalization data, the best clustering results are achieved with the use of 3 clusters, as the clustering error is minimal.

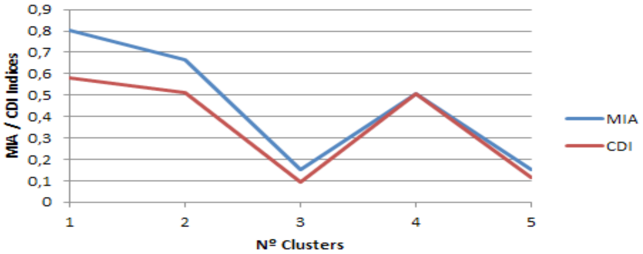

Figure 7: MIA and CDI results for average normalization values for 24 hour period

Much like the results with regular data, when using 2 clusters, a clear separation of residential houses and buildings from commercial buildings is visible. Again the two commercial buildings (corresponding to loads 1 and 2) have been allocated to cluster 1, and the rest of the loads, corresponding to residential consumers, have been aggregated in cluster 2. This can be observed in Figure 8 which presents the load profiles of consumers that have been grouped in cluster 1 and in cluster 2 using the average normalization data.

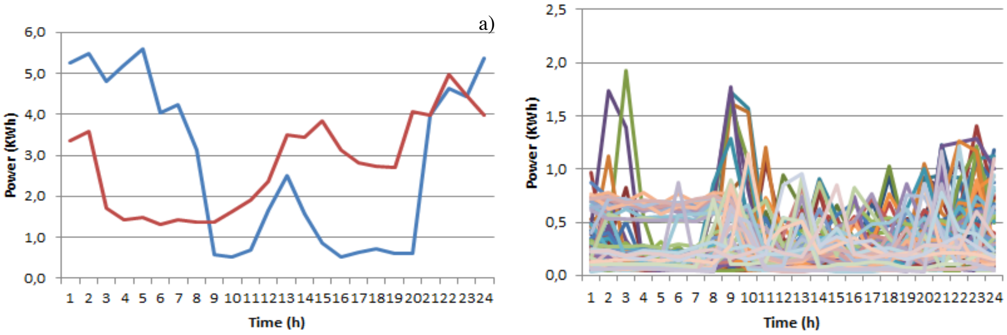

Figure 8: Consumption profile of load allocated to: a) cluster 1; b) cluster 2

From Figure 8 it is visible that cluster 1 includes the two commercial buildings, with very distinct load profiles, and cluster 2 includes all the residential buildings and houses.

The similarity of results, when comparing with the non-normalized data, is related to the fact that we are working with data that are not normalized. As said before, the reduction of the data to the same scale is fundamental because certain non-parametric algorithms assume implicitly that the distances in different directions of the input space have the same weight. Without recourse to normalization variables with large numerical values, the non-normalized data, can dominate the effects of variables with smaller but equally important values.

5.4. Difference process

In this process, the value of the microproduction, generated in the bus associated to the same load, was subtracted from the value of the consumption of that load. This calculation is performed for loads from 1 to 17, these are the loads, which are on buses with associated microproduction. MIA and CDI are used to analyse the clustering error, they are presented in Figure 9.

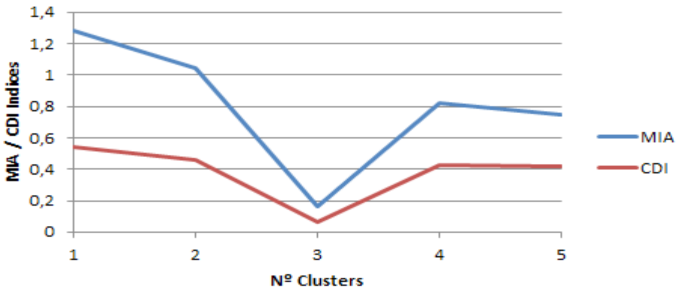

Figure 9: MIA and CDI results for difference normalization values for 24 hour period

In difference normalization process, the best clustering results are achieved with the use of 3 clusters, as the clustering error is minimal, similar to what happened in previous cases. When using 2 clusters, a clear separation of residential houses and buildings from commercial buildings is visible. It is also visible that the two commercial buildings (corresponding to loads 1 and 2) have been allocated to cluster 1, and the rest of the loads, corresponding to residential consumers, have been aggregated in cluster 2.

As we can see in Figure 10, when considering the grouping process with 3 clusters, the difference is that there is a even better separation of consumers types, commercial buildings were allocated to cluster 1, residential houses to cluster 2 and residential building to cluster 3. In this case the two types of commercial buildings stayed in the same cluster, although they present very different load profiles.

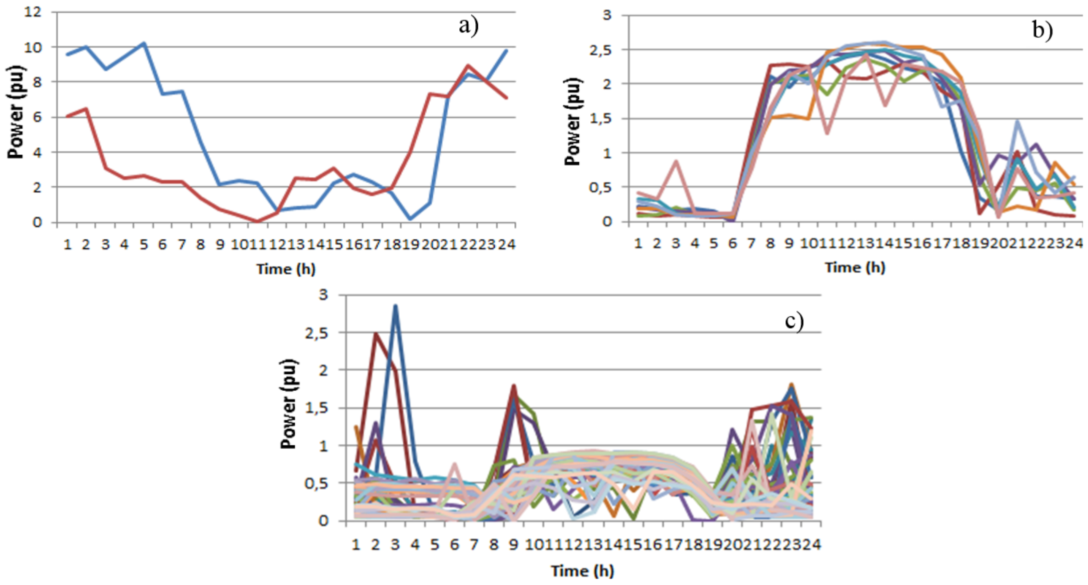

Figure 10: Consumption profile of load allocated to: a) cluster 1; b) cluster 2; c) cluster 3

5.3. Customized Normalization

The clustering process is performed for different numbers of clusters (from 2 to 6). MIA and CDI are used to analyse the clustering error. Figure 3 presents the comparison of the MIA and CDI error values that are achieved when using from 2 to 6 clusters, with each of the three considered methods: non-normalized data, normalized data using the regular method, and customized normalization method.

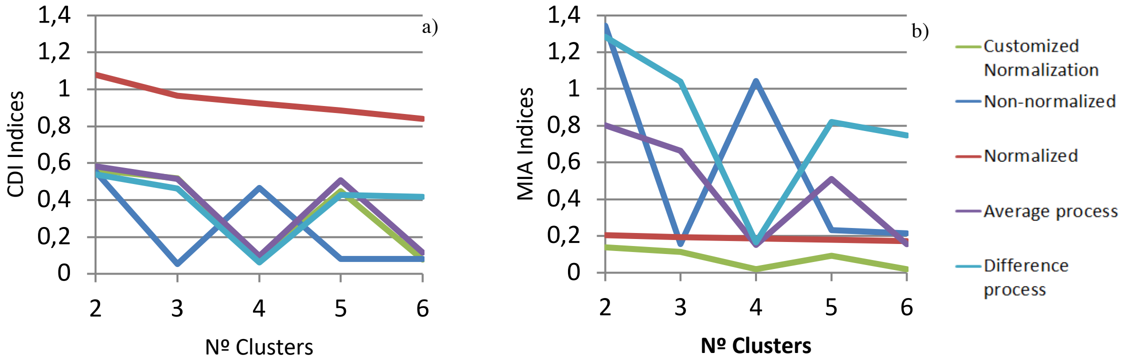

Figure 11: CDI a) and MIA b) results for different numbers of clusters, using non-normalized, normalized, average and difference process and customized normalization

In Figure 11 it is possible to see that the customized normalization has a lower error when compared with the previous methods. The best clustering results are achieved with the use of 4 clusters, as the clustering error is minimal. With the use of 4 clusters, the two commercial buildings are separated into a different cluster each, the third cluster allocates some of the residential buildings that have similar load profiles, and the rest of the loads, corresponding to residential houses and some residential buildings, have been aggregated in the final cluster. Figure 12 represents the load profile of the different consumer types considering 3 clusters. Figure 12a) represents the consumption profiles of the two commercial buildings, which, as it is possible to see, have very different consumption profiles, especially during the night. This is why they were allocated into different clusters when 4 clusters are considered. In Figure 12b) and 12c) represent the consumption profiles of the loads allocated to the other two clusters.

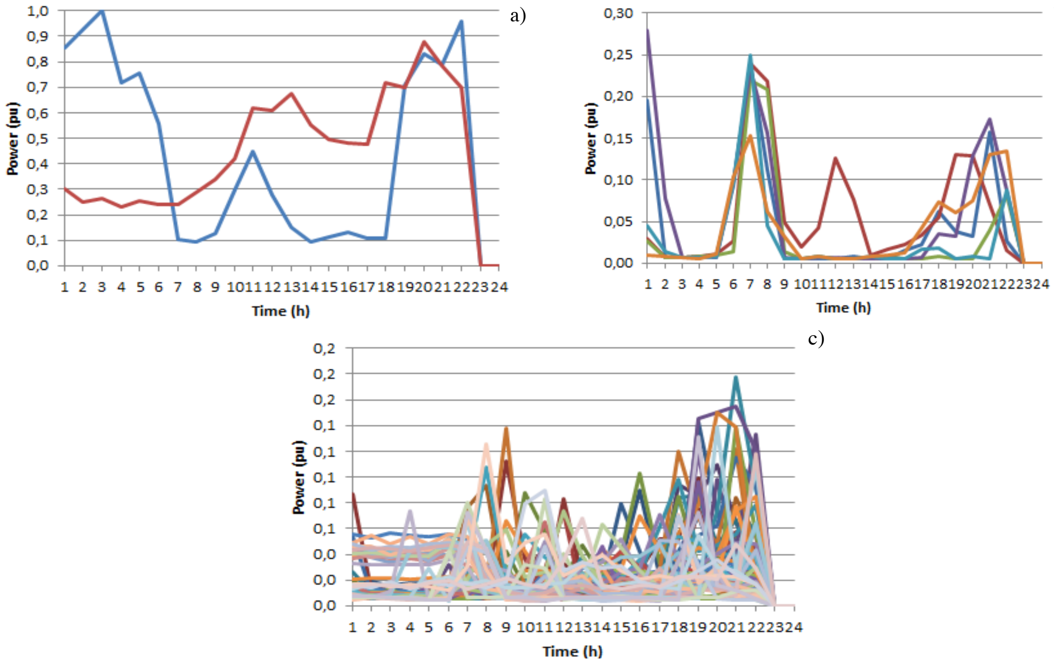

Figure 12: Consumption profile of: a) commercial building cluster 1; b) residential building cluster 2; c) residential house cluster 3

From Figure 12 the separation of the load profiles is evident. It is also visible that a separation taking into account the gross amount of consumption of each consumer has been accomplished, as commercial consumers, which present much higher consumption values, have been separated from the residential consumers. The proposed customized normalization brings, therefore, clear advantages to the RemT tariff definition process. It enables to clearly identify different consumers, taking into account their consumption tendency and amount, therefore breaking the way for an objective and fare definition of dynamic electricity tariffs, which can suitably fit each of the identified groups, i.e. consumers with similar consumption tendencies, taking into account their dimension. The new type of normalization, when compared with the previous normalization types, allows an even more clear separation of consumer types, which is evident from the load profile graphs that show the separation into different clusters, and also by the MIA and CDI values, which show that the proposed method achieves smaller clustering error values than the other methods.

6. Conclusion

This case study demonstrated the usefulness and advantage of data mining methodologies, based on the application of clustering process to group typical load profiles of consumers according to their similarity to allow proposing specific consumption tariffs to each group, so that consumers load profile is taken into account to meet the objectives of the SG aggregator. This work allows the development of a tool that provides a decision support for VPP definition of best tariff and remuneration to apply to each aggregated player, RemT. To develop RemT a clustering methodology that uses different data normalization methods was presented, and a new customized normalization method has been introduced.

Currently there is a gap in what concerns VPP aggregated players' tariffs and remuneration. In order to overcome this problem, developing appropriate methods is essential.

The results of the presented case study, based on real consumption data, show that the customized normalization method combines the advantages of both previous approaches, so as to achieve more consumer groups that capture both differences in the quantities of consumption, as well as the trends of consumer profiles along hours. This is crucial, according to different consumers' types and profiles, as it works as a basis for personalized and dynamic consumption tariff definition. Thus normalization is the same made between 0 and 1, but there is visible difference between higher and lower consumption among different consumers, and the evolution of consumption of each consumer profile is also captured. RemT mechanism is evolving to become a crucial tool to go a step forward in electricity markets simulation, by enabling a fair and dynamic means to define electricity tariffs for different types of consumers.

7. References

Anil K. Jain, 2010. "Data Clustering: 50 years beyond K-Means". Pattern Recognition Letters, Elsevier, Vol. 31, Issue 8, pp.651-666.

Anil K. Jain, M.N Murty, P.J. Flynn, 1999. "Data Clustering: A Review." ACM Computing Surveys, 31 (3). pp. 264-323.

Anthony C. Chrysopoulos, Andreas L. Symeonidis, Pericles A. Mitkas, 2009. "Improving agent bidding in power stock markets through a data mining enhanced agent platform". Agents and Data Mining Interaction"; Lecture Notes in Computer Science

Blumsack S and Fernandez A., 2012. "Ready or not, here comes the smart grid!" Energy. 2012; 37(1): 61-8

CAISO – California Independent System Operator. Available: http://www.caiso.com [accessed on July 2017]

Canizes B., Silva M., Faria P., Ramos S., Vale Z., 2015. "Resource Scheduling in Residential Microgrids Considering Energy Selling to External Players", Power Systems Conference (PSC 2015), South Carolina, USA, 10-13 March

C. Ribeiro, T. Pinto, Z. Vale, 2016. "Customized Normalization Method to enhance the Clustering process of Consumption Profiles", 7th International Symposium on Ambient Intelligence (ISAMI'16), Spain, 1st-3rd June

C. Ribeiro, T. Pinto, M. Silva, S. Ramos, Z. Vale, 2015. "Data Mining approach for Decision Support in real data based Smart Grid scenario" IATEM, Valencia, Spain, 1-4 September

C. Ribeiro, T. pinto, H. Morais, Z. Vale, G. Santos, 2013. "Intelligent Remuneration and Tariffs in for Virtual Power Players", IEEE PowerTech (POWERTECH) Grenoble, France, 16-20 June

C. KienY, B. Berseneff, N. Hadjsaid, Y. Besanger, J. Maire, 2009. "On the concept and the interest of Virtual Power plant: some results from the European project FENIX", IEEE Power and Energy Society General Meeting, Calgary, Canada, 26-30, July

Dore A. and Regazzoni C., 2010, "Interaction Analysis with a Bayesian Trajectory Model", IEEE Intelligent Systems, vol. 25, no. 3, pp. 32–40

EPEXSPOT – European Power Exchange Products Day-Ahead Auction, 2015. Available: https://www.epexspot.com/en/product-info/auction, [accessed on July 2017].

Erev, I. and Roth, A.,1998. "Predicting how people play games with unique, mixed-strategy equilibria", American Economic Review, vol. 88, pp. 848–881

G. Chicco and I. Ilie, 2009. "Support Vector Clustering of Electrical Load Pattern Data". IEEE Transactions on Power Systems, vol.24, no.3, pp.1619-1628, August

G. Chicco, R. Napoli, P. Postolache, M. Scutariu, C. Toader, 2003. "Customer Characterization Options for Improving the Tariff Offer", IEEE Transactions on Power Systems, Vol. 18, nº 1, pp. 381-387, February

G. Gan, C. Ma, J. Wu, 2007. "Data Clustering Theory, Algorithms and Applications", ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA

I. Praça, C. Ramos, Z. Vale, M. Cordeiro, 2003. "MASCEM: A Multi-Agent System that Simulates Competitive Electricity Markets", IEEE Intelligent Systems, 18, 6, pp. 54-60

MIBEL - Mercado Ibérico de Electricidade, 2017. Available: http://www.mibel.com/, [accessed on July 2017]

Mohammad Shahidehpour, Hatim Yamin, Zuyi Li., 2002. "Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management", Wiley-IEEE Press, pp. 233-274

Nord Pool Spot - Trading, Day-ahead market Elspot, 2016. Available: http://www.nordpoolspot.com/TAS/Day-ahead-market-Elspot/, [accessed on July 2017].

Oliveira P., Pinto T., Morais H., Vale Z., 2012. "MASGriP - A Multi-Agent Smart Grid Simulation Plataform," IEEE Power and Energy Society General Meeting, San Diego, California USA, pp. 1-10

S. Ramos, J. Duarte, J. Soares, Z. vale, F. Duarte, 2012. "Typical Load Profiles in the Smart Grid Context – A Clustering Methods Comparison", IEEE Power and Energy Society General Meeting, San Diego CA, USA, 22–26 July

Sharma K.C., Bhakar R., Tiwari, H.P., 2014. "Strategic bidding for wind power producers in electricity markets." Energy Conversion and Management 2014, 86, 259–267, DOI: 10.1016/j.enconman.2014.05.002.

Sioshansi, F.P., 2013. "Evolution of Global Electricity Markets – New paradigms, new challenges, new approaches", Academic Press.

Sousa T., Morais H., Vale Z., Faria P., Soares J., 2012. "Intelligent Energy Resource Management Considering Vehicle-to-Grid: A Simulated Annealing Approach," Smart Grid, IEEE Trans. on, 3, 535-542.

T. Pinto, Z. Vale, F. Rodrigues, H. Norais, I. Praça, 2011. "Strategic Bidding Methodology for Electricity Markets using Adaptive Learning", Modern Approaches in Applied Intelligence, vol. 6704, pp.490-500

T. Pinto, Z. Vale, H. Morais, I.Praça, C. Ramos, 2009. "Multi-Agent Based Electricity Market Simulator With VPP: Conceptual and Implementation Issues", IEEE PES General Meeting.

Z. Vale, T. Pinto, I. Praça, H. Morais, 2011. "MASCEM - Electricity markets simulation with strategically acting players", IEEE Intelligent Systems, vol. 26, n. 2, Special Issue on AI in Power Systems and Energy Markets.

Z. Vale, H. Morais, P. Faria, H. Khodr, J. Ferreira, P. Kadar, 2010. "Distributed Energy Resources Management with Cyber-Physical SCADA in the Context of Future Smart Grids", 15thIEEE Mediterranean Elect. Conf., Malta, 25-28 April.