sustainability

MDPI

Article

Sustainable Development with Smart Meter Data Analytics Using NoSQL and Self-Organizing Maps

Simona-Vasilica Oprea [1,*], Adela Bâra [1], Bogdan George Tudorică [2] and Gabriela Dobrița (Ene) [1]

1 Department of Economic Informatics and Cybernetics, The Bucharest University of Economic Studies, Piaţa Romană square, 010374 Bucharest, Romania; bara.adela@ie.ase.ro (A.B.); gabrielaene02@gmail.com (G.D.)
2 Department of Cybernetics, Economic Informatics, Finance and Accountancy, Petroleum-Gas University of Ploiești, Bucureşti avenue, 100279 Ploiești, Romania; tudorica_bogdan@yahoo.com
* Correspondence: simona.oprea@csie.ase.ro

Received: 7 April 2020; Accepted: 22 April 2020; Published: 23 April 2020

Abstract: The smart metered electricity consumption data and high dimensional questionnaires provide useful information for designing the tariffs aimed at reducing electricity consumption and peak. The volume of data generated by smart meters for a sample of around four thousand residential consumers requires Not only Structured Query Language (NoSQL) solutions, data management and artificial neural network clustering algorithms, such as Self-Organizing Maps. In this paper, we propose a novel methodology that handles a large volume of data and extracts information from electricity consumption measured at 30 min and from complex questionnaires. Five three-level Time-of-Use tariffs are altered and investigated to minimize the consumers' payment. Then, input data analysis revealed that the peak consumption is influenced by a segment of consumers that can be targeted to flatten the peak. Based on simulations, more than 23% of the peak consumption can be reduced by shifting it from peak to off-peak hours.

Keywords: NoSQL; self-organizing maps; smart meters; electricity consumption; questionnaire

1. Introduction

From vertically integrated utility companies to the distributed energy sources, this industry has transformed tremendously, facing new challenges that come along with the Information Communication Technology (ICT) progress. Gradually, the energy generated in remote large power plants has been replaced by new distributed generation sources located close to residential areas, which modified the unidirectional flows direction. Now, the electricity flows in both directions, from and to the grid, and it is frequently measured by Smart Meters (SM) that generate a large volume of data. Also, the consumers own more and more modern devices with IoT connectivity [1] that can be remotely controlled to meet certain objective functions. During trial periods, the electricity consumers are subject to complex questionnaires (pre- and post-trial) that can be deployed by regulators, grid operators or suppliers, to better understand and predict consumers behavior and trends. Their answers are useful in designing the demand response strategies, including advanced Time-of-Use (ToU) tariffs [2], implemented thanks to communications progress and SM that can incentive consumers to energize their appliances at lower tariffs. However, the implementation of SM is expensive and require trials aiming to identify the opportunity to install SM at large scale [3]. Hence, the consumption, appliances data and consumer answers (opinions) are significant data sources, pouring from high-dimensional surveys and SM, which will be analyzed in this paper.

In this paper, we propose a methodology of analyzing large volumes of data aiming to identify electricity consumers' behavior and their potential to reschedule the appliances.

The paper is structured in seven sections. The first one, Introduction, is a short description of data sources and the context of our research, dealing with large volume, electricity consumption and high-dimensional questionnaire data. The second section, Literature review, presents several scientific researches that treat similar topics. The third section, Input Data, deals with the source, structure, formats and specificity of the data. It includes metering and questionnaire data processing, indicating the various tools and techniques we used to prepare and analyze the data. The fourth section, Methodology, describes the scientific methodology for analyzing the questionnaire data. The fifth section, Results, includes the findings, grouped by the origin of data we draw conclusions on. The sixth section, Discussion, includes a comparative table that analyses several research papers and the current study from the results point of view. Finally, the last section comprises the conclusion for the entire paper, as per usual.

2. Literature Review

2.1. Self-Organizing Maps

Artificial Neural Networks were introduced in 1943 by Warren McCulloch and Walter Pitts. A very simple model of the biological neuron was proposed in [4]: it has one or more binary inputs and only one binary output. The activation function describes the rules used for determining the output. The input layer corresponds to the data received from the interaction with the external world, the hidden layer contains the executed computation based on the provided functions, and the output layer corresponds to the information received. Inside the network, all the information is propagated from one layer to another, until the output is obtained. Knowledge is represented in synaptic weights between the neurons [5].

Artificial intelligence techniques, along with statistical methodologies and data visualization methods, represent one main groups of classification techniques used for data mining. In terms of supervised learning, training data consists of specific input–output mapping examples that the neural network must approximate, while, in unsupervised learning, the neural network must discover significant patterns on the features of data through unlabeled training samples. The unsupervised learning approach may be implemented from two perspectives: self-organized learning (following neurobiological structures) or statistical learning theory (traditionally used in machine learning).

The self-organized learning algorithm requires a competitive learning rule based on which the neurons in the competitive layer compete for the opportunity to interact with data features. The neuron with the highest total turns on by adopting the *winner-takes-it-all* strategy. The algorithm contains a set of rules that define the local behavior for a specific neuron—the sum of adjustments that are made to the synaptic weights are contained in the local neighborhood of the neuron. D. Hebb was the first one to propose the following rule as the basis of associative learning, widely known as Hebbian learning: when an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic changes take place in one or both cells, such that A's efficiency as one of the cells firing B is increased [6].

The most common unsupervised learning tasks include clustering, dimensionality reduction, anomaly detection or density estimation. Clustering is one of the unsupervised learning techniques used for structure discovery in data, and the *k-means* algorithm is undoubtedly one of the most popular clustering algorithms, but it also has limitations: firstly, knowing the number of clusters might be a problem, and the algorithm does not behave very well for clusters that have various sizes, different densities or non-spherical forms.

Kohonen demonstrated for the first time that *Self Organizing Maps* (SOM) algorithm can be implemented for any data set for which a degree of dissimilarity or similarity of data is known [7]. K-means clustering is similar to the Kohonen algorithm for neighborhoods with the size equal to one, and for larger neighborhoods it is a generalization of k-means, in which each weight adapts toward the center of its cluster of patterns and its neighbors' clusters, resulting in an ordered mapping that tends to preserve the topological structure of the input distribution [8,9].

Competitive learning is a particular class of unsupervised learning algorithms, in which the neurons compete to be activated, with the result that only one will be activated at a time. To achieve this, a feature map or a representation of the training data is needed.

2.2. Large Volume of Data and Electricity Consumers' Behavior

A comparison between big data frameworks and mining algorithms for the MapReduce (MR) solution is extensively performed in [10]. The authors treat the noise, outliers, incomplete and inconsistent data, bottlenecks on data mining algorithms, security and privacy issues that come with big data.

He Y. et al. [11] identifies the challenges in clustering a large size of data sets, and they design a parallel density-based algorithm for clustering developed with a four-stage MR method, adopting a quick partitioning of large scale non-indexed data sets. A DBSCAN clustering algorithm using the complex big data processing solution Spark is proposed in [12], efficiently evaluating the scalability of the algorithm with a various number of processing cores.

Numerous studies that assessed the effects of installing real-time display of electricity consumption have been developed. It has been determined that this single action can generate a small reduction in the consumption level, around 3–5%, together with a reduction of the consumption peak, but it did not lead to lower carbon or greenhouse gas emissions [13]. Additionally, Alahmad M.A. et al. studies the result of installing different types of devices to monitor real-time energy consumption in residential houses, where around 41% of actual energy consumption is wasted. The aggregated data after 30 days showed an insignificant drop in energy consumption [14].

The main subject in [15] is the consumer's extremely complex behavior, as one's actions and options would be often unpredictable. It was proven that consumers are tempted to do social comparisons, respect social constrains and norms, follow the behavioral patterns of others and reluctantly change old electrical appliances with high energy consumption. Still, the financial incentives for modernizing appliances, as well as social influence, can motivate consumers. Another study [16] seeks to analyze by what degree the consumers can be activated to change their behavior. Therefore, the service designers have to interact with the consumers and foresee the implications, from both a technical, as well as a financial point of view. Additionally, smart grid solutions are investigated in [17] to find a better technique to balance and maintain equilibrium between generation and consumption, and thus achieve maximal efficiency. The effects on consumption for the people that use more performant feedback technologies, providing real-time information that is accessible through mobile devices, are revealed in [18]. It was found that a 5.7% average reduction in consumption is obtained, especially at peak hours, but for a short timeframe (only 8 weeks).

Kendel A. et al. studies business models for SM in France and, connected to them, possible incentive systems, smart tariffs and other instruments aimed at changing household consumers' behavior, integrating SM with solutions capable of exploiting renewable energy sources and energy stocks [19]. As new IT&C technologies keep developing, we can already talk about smart homes integrating smart appliances and meters, home gateways, communication systems, sensors, controllers, energy management systems, dynamic tariff systems and artificial intelligence predictions [20].

The residential consumers' behavior is also studied in [21], identifying the common consumption practices and uncovering the ways to influence the financial and non-financial factors [22]. To generate actions that would reduce consumption, the Japanese study shows that households responded well to financial incentives and less to the non-financial ones. Another study from Austria on 1500 households analyzes the effects of real-time feedback provided by the consumers, together with a set of information regarding energy saving measures that lead to an average of approx. 4.5% economy in electricity consumption [23].

Additionally, in [24], the authors studied the effects of smart display for households and show that, even if households received educational materials, legislative framework information regarding these issues and their short-term consumption dropped by 9%, which was not enough for long-term. Hence, the effect of introducing ToU tariffs for residential consumers was investigated for a northern Italian county [25], demonstrating that even if the consumption peak in the morning was diminished and the bills were reduced by 2.21%, the average electricity consumption increased by 13.69%. Under these circumstances, other researchers studied the consumption behavior in 2000 households in Sweden for 4 years, to identify measures and causes that can influence the reduction of residential consumption [26]. They observe that those consumers that had a web available consumption display achieved energy savings around 15%, as the information led to action.

Usually, complex surveys offer extensive knowledge about their respondents, while simultaneously presenting a challenge due to their high dimensionality and due to the occurrence of variables of different types. According to [27], a data set is considered high-dimensional if it exceeds 16 variables. Therefore, with more than 140 questions in the pre-trial questionnaire, and over 190 questions in the post-trial questionnaire that we analyze, both easily fall into this category.

3. Input Data

3.1. Advanced Electricity Tariffs

The Commission for Energy Regulation from Ireland administrated a project to identify the electricity consumers behavior regarding peak and overall electricity usage considering the implementation of SM technology, ToU tariffs and other stimuli [3].

Three types of consumers—*small and medium enterprises* (SMEs), residential consumers and others were involved in a trial period in 2010. It was preceded by a pre-trial that consists in designing ToU tariffs and *demand side management* (DSM) incentives, such as: recruitment and grouping the consumers, consumption monitoring and reduction incentive, recruitment, tariff allocation matrix and pre-questionnaire, and was followed by an account reconciliation and post-questionnaire. The consumers' bills from the test groups were regulated by applying the flat tariff instead of the trial ToU tariffs. In this paper, we concentrated our analyses on residential consumers, because different DSM strategies were envisioned for SMEs and others.

The test groups that each corresponds to a ToU tariff were set to DSM, while the control group was billed as before with the flat tariff (tariff E equals 14.1c/kWh). Four three-level ToU tariffs and one tariff for weekend (W), as in Figure 1, were created and applied for residential consumers.

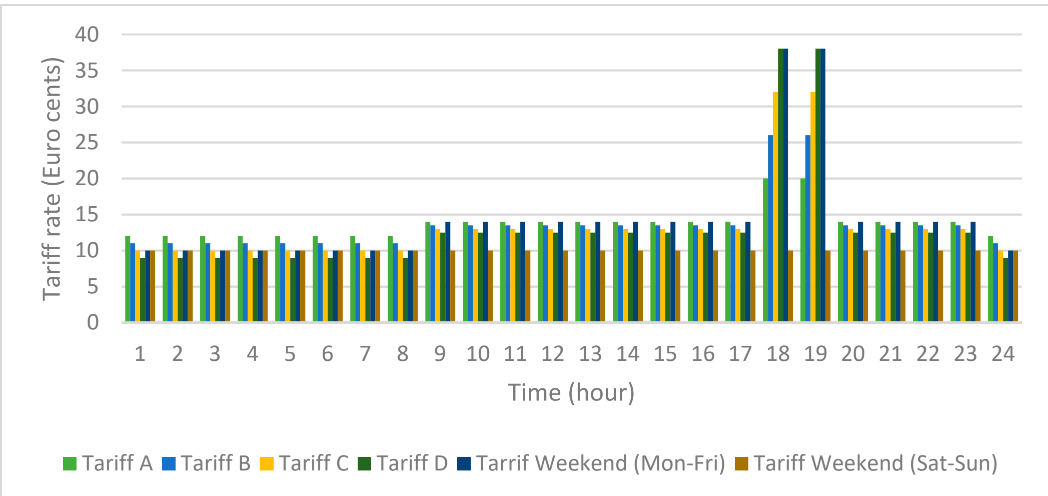

Figure 1. Time-of-Use (ToU) tariffs.

C and D are remarkable by higher peak rates and lower off-peak rates, also W tariff has a higher peak rate from Monday to Friday, but constant low rate during weekend days. Such tariffs reward consumers that shift their consumption from peak to off-peak intervals. On the other hand, A and B are less rewarding, but their peak rates are also less penalizing.

3.2. Brief on Pre- and Post-Trial Questionnaires

The pre-trial questionnaire consists in 143 questions, and provides data about respondents from different points of view, such as: sex, age, employment status, education, income, Internet access, broadband, usage, size and structure of the family, positive or negative attitude to reduce or shift the electricity usage, household description (surface, no of bedrooms, construction year), actual heating system, appliances, measures to reduce the electricity usage and expectations from the trial period.

Figures 2–8 provide insights regarding the consumers attitude before trial. Figure 2 synthesizes the data gathered through the pre-trial questionnaire regarding sex and age categories that benefit or not from the Internet access.

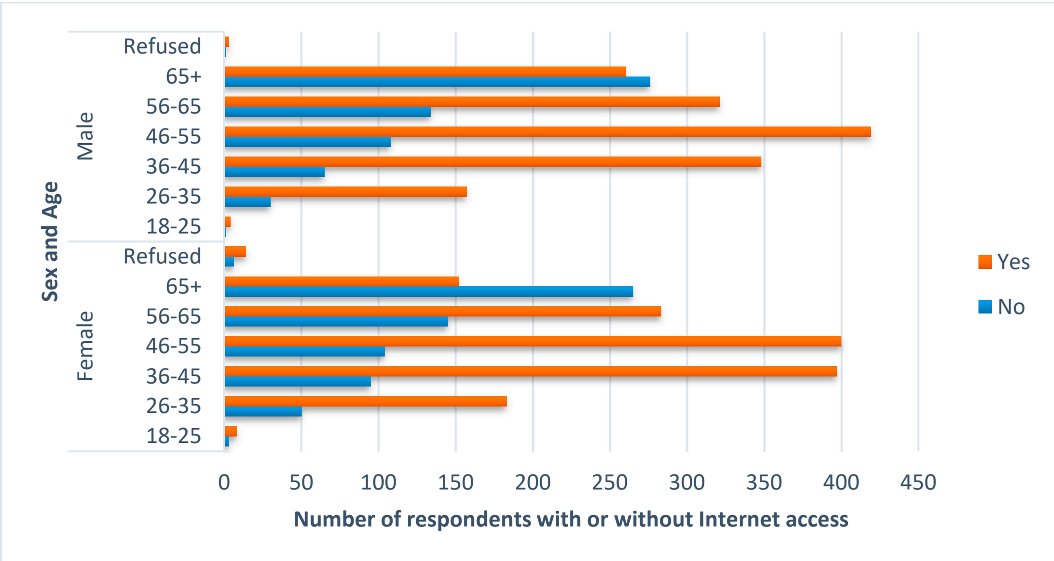

Figure 2. Pre-trial respondents' profiles from the Internet access point of view.

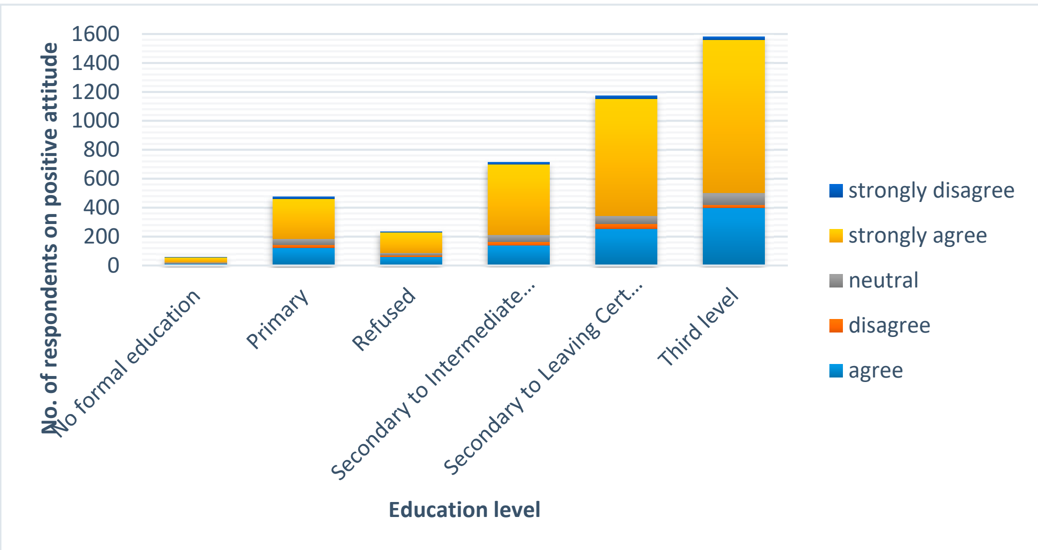

Figure 3. Pre-trial respondents' positive attitude to reduction of electricity usage.

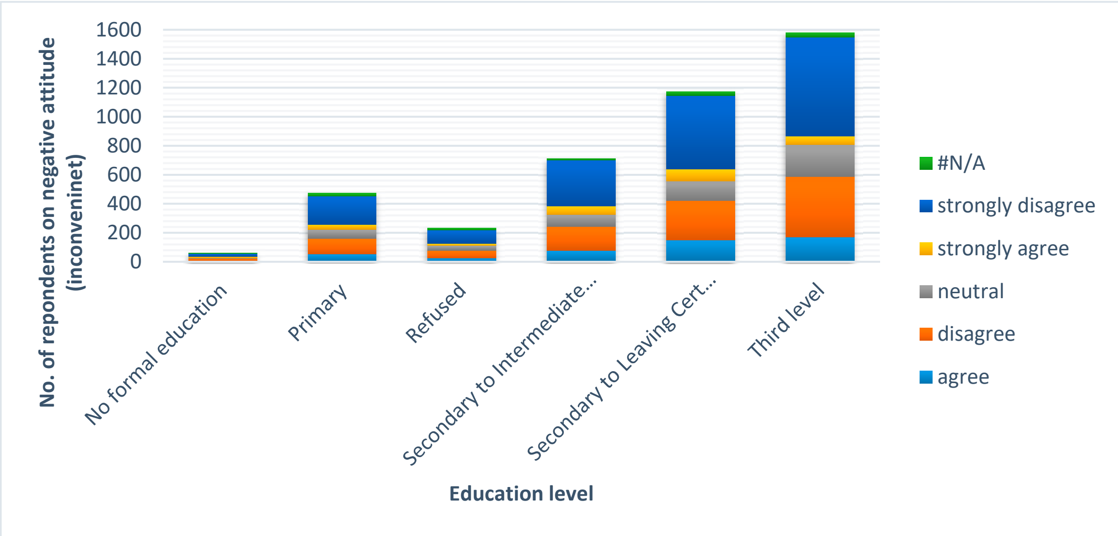

Figure 4. Pre-trial respondents' negative attitude to reduction of electricity usage.

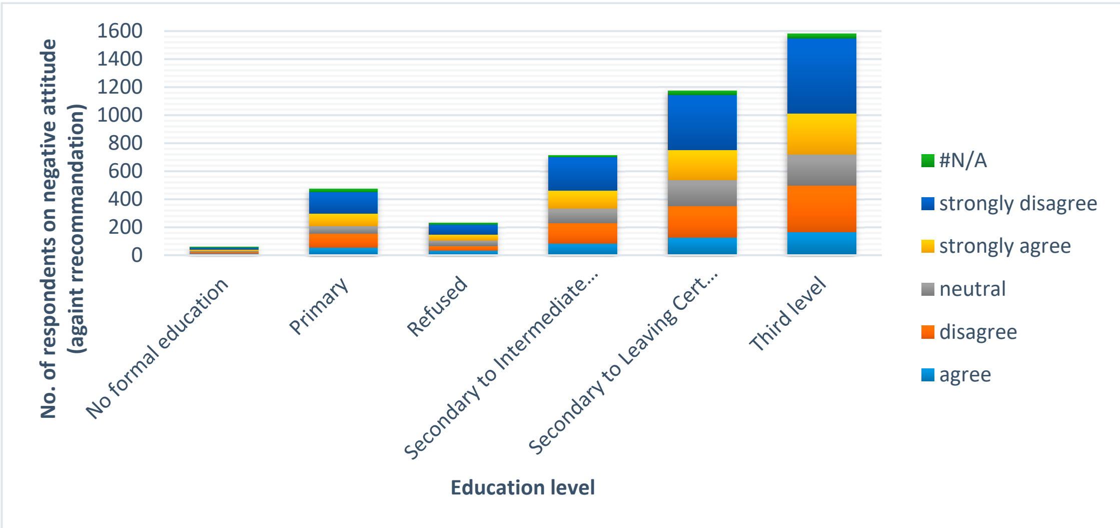

Figure 5. Pre-trial respondents' negative attitude to the electricity usage recommendation.

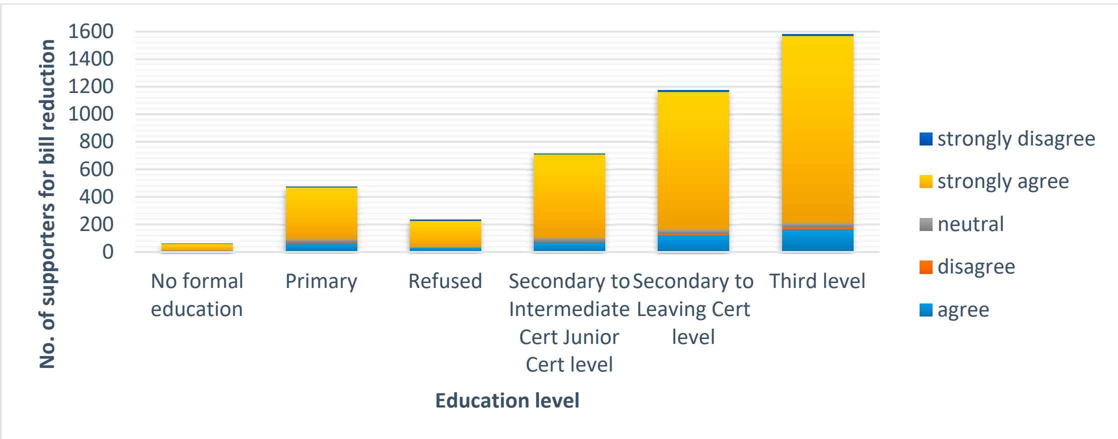

Figure 6. Pre-trial respondents' electricity usage motivation regarding bill reduction.

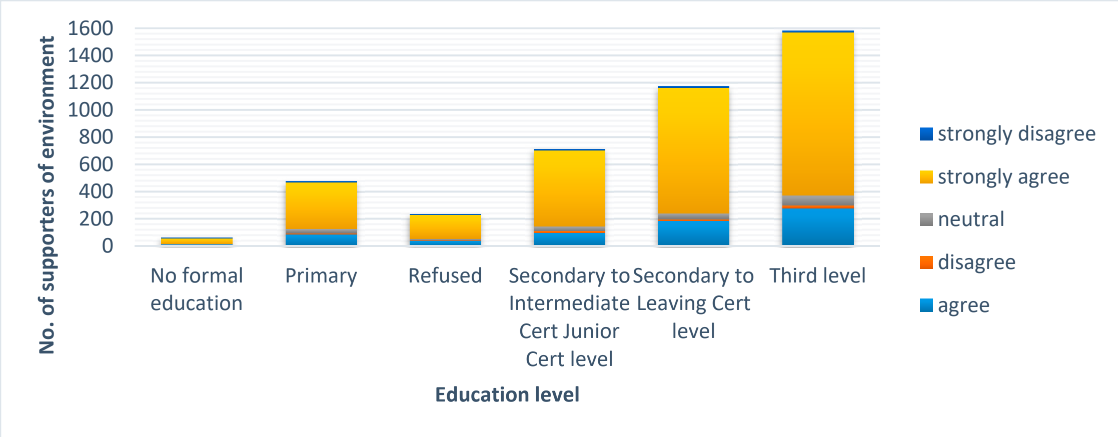

Figure 7. Pre-trial respondents' electricity usage motivation regarding environment care.

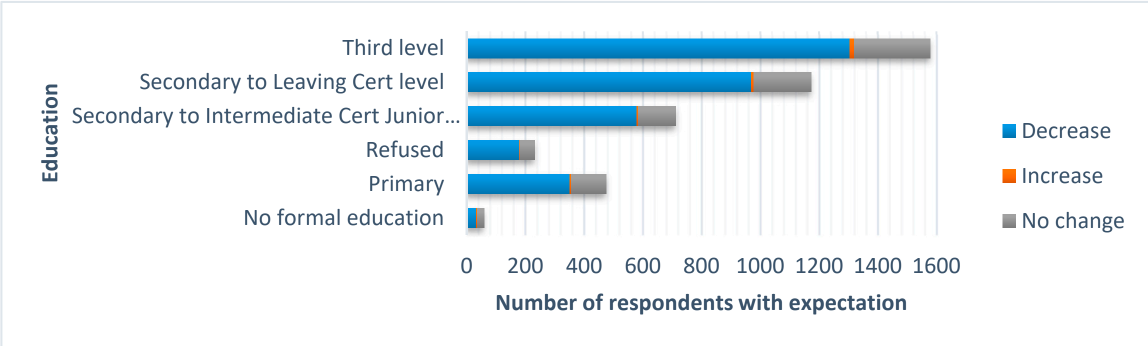

Figure 8. Pre-trial respondents' expectation regarding electricity bill.

Figure 3 shows the attitude of the respondents on education levels that reacted to the statement "*I/we would like to do more to reduce electricity usage*". Most of them strongly agreed or agreed. Similar results are obtained when the respondents were grouped on employment status or age.

Figure 4 shows the attitude of the respondents on education levels that reacted to the statement "*It is too inconvenient to reduce our usage of electricity*".

Figure 5 shows the attitude of the respondents on education levels that reacted to the statement "*I do not want to be told how much electricity I can use*". The answers were less polarized, whereas the proportions of the respondents that strongly disagreed or strongly agreed are not very different.

Figure 6 shows the attitude of the respondents on education levels that reacted to the statement "*I/we am/are interested in changing the way I/we use electricity if it reduces the bill*". In this case, the answers were well polarized around "strongly agree" and "agree".

Figure 7 shows the attitude of the respondents on education levels that reacted to the statement "*I/we am/are interested in changing the way I/we use electricity if it helps the environment*". In this case, the answers were also well polarized around "strongly agree" and "agree". Only the neutral answers are more often than in the previous figure.

Figure 8 shows the expectation of the respondents on education levels that reacted to the question "*How do you think that your electricity bills will change as part of the trial?*". Most of the respondents (81%) expected to decrease the electricity bill, while 18% answered with "No change".

The post-trial questionnaire consists of 234 questions answered by the consumers who finalized the trial period. They measure the attitude adjustments in terms of electricity usage to the pre-trial findings. The majority of questions referred to the consumers' opinions regarding the perceived electricity usage and the impact the ToU tariffs and other DSM stimuli.

Figures 9–16 provide insights regarding consumers attitude after trial. Figure 9 shows the attitude of the respondents on age categories that reacted post-trail to the statement *"I/we would like to do more to reduce electricity usage"*. Most of them strongly agreed or agreed.

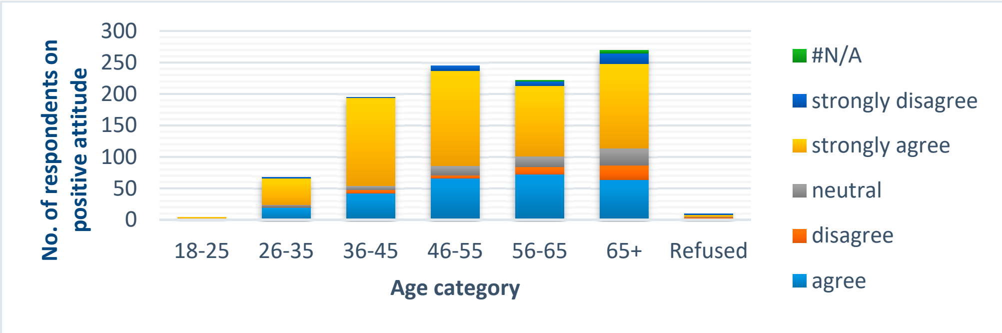

Figure 9. Post-trial respondents' positive attitude to reduction of electricity usage.

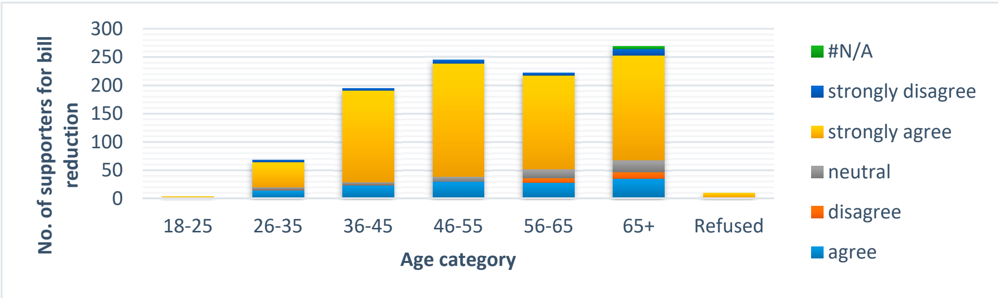

Figure 10. Post-trial respondents' electricity usage motivation regarding bill reduction.

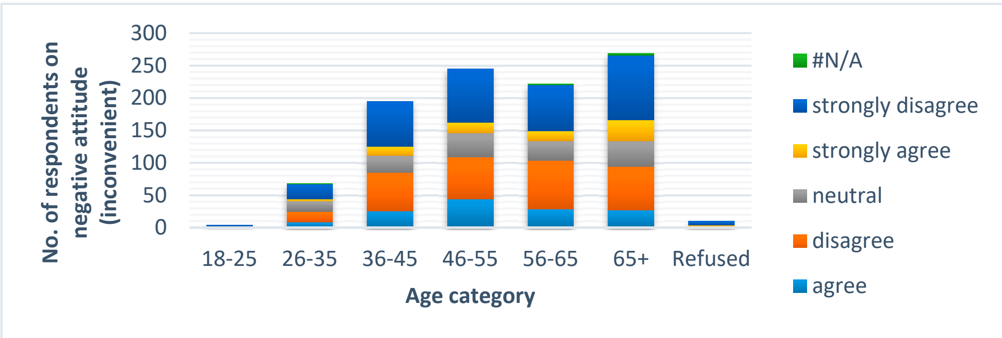

Figure 11. Post-trial respondents' negative attitude to reduction of electricity usage.

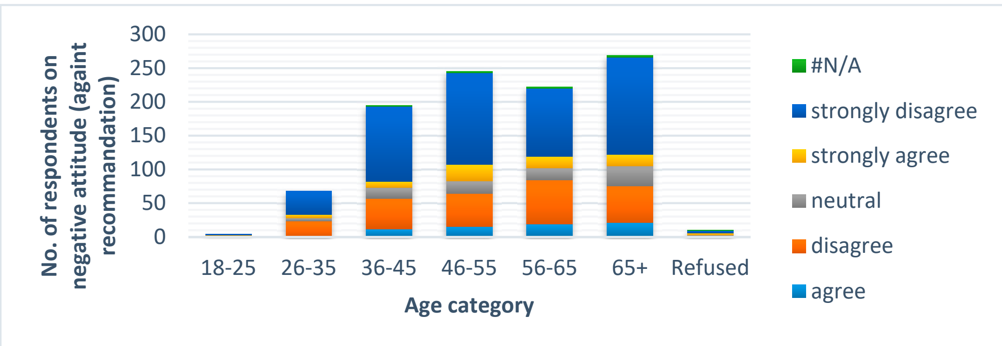

Figure 12. Post-trial respondents' negative attitude to the electricity usage recommendation.

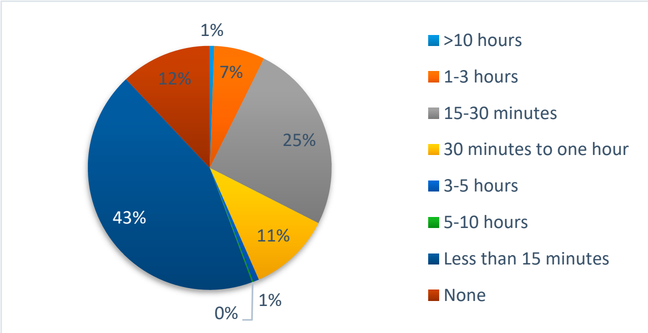

Figure 13. Post-trial respondents' time spent to understand the tariffs.

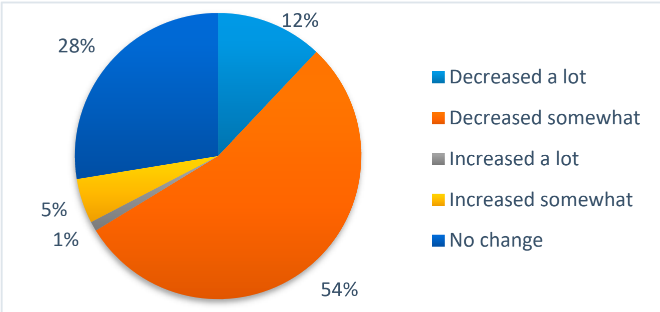

Figure 14. Post-trial respondents' answer regarding electricity bill change.

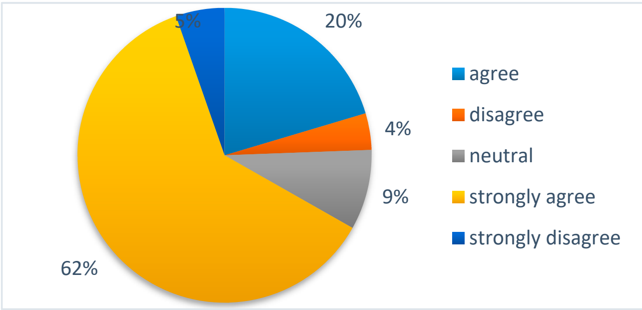

Figure 15. Post-trail tariff perception regarding behavioral changes.

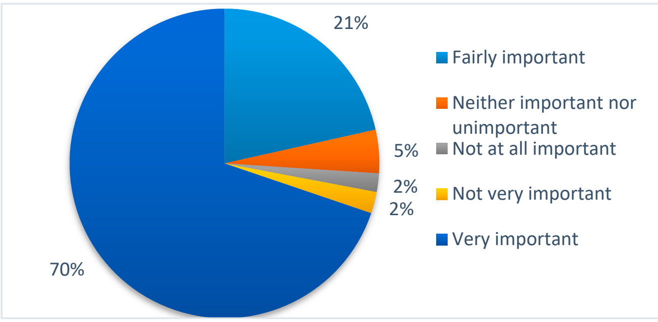

Figure 16. Impact of the ToU tariffs as perceived by post-trial respondents.

Figure 10 shows the attitude of the respondents on age categories that reacted post-trail to the statement *"I am interested in changing the way I use electricity if it reduces the electricity bill"*. Most of them strongly agreed or agreed.

The attitude of the respondents on age categories that reacted to the statement *"I am interested in changing the way I use electricity if it helps the environment"* is also similar.

Figure 11 shows the attitude of the respondents on education levels that reacted to the statement *"It is too inconvenient to reduce our usage of electricity"*.

Figure 12 shows the attitude of the respondents on education levels that reacted to the statement *"I do not want to be told how much electricity I can use"*.

Figure 13 shows the time the respondents spent to study and understand the tariff structure. A total of 43% of the respondents spent less than 15 min, while 25% of the respondents spent between 15 and 30 min. Similar proportions were recorded in case of the time spent by the respondents to understand the overall load reduction, electricity monitor and electricity usage statement as part of DSM measures.

Figure 14 shows the answer of the respondents to the question *"By what amount do you think that your electricity bills changed as a result of the trial?"* A sum of 66% of the respondents considered that their electricity bills decreased somewhat or a lot, while 29% of the respondents considered that their electricity bills had not changed. Similar results were obtained when the respondents evaluated the change of the electricity usage or electricity usage at peak hours.

A total of 82% of the respondents strongly agreed or agreed that the tariff helped them change consumption behavior, as shown in Figure 15.

The encouragement perceived by respondents to reduce the electricity usage offered by tariff structure is given in Figure 16.

Similar results were obtained in case of the encouragement perceived by the respondents to reduce the electricity usage by electricity monitor and additional information from the invoice.

3.3. Consumption Data Processing

The available consumption data consisted of text data files, with three fields being the customer identifier, a compound field containing date and time information and consumption data corresponding to 30-minute samples.

The data files contained 157,992,996 records regarding 12 months of metering for 6435 customers (4225 residential customers, 485 small and medium enterprises and 1725 of another type).

MongoDB was chosen to be used as a data storage solution. Combinations of various languages, techniques and software applications were used, in order to maximize data analysis capabilities (the C#, R, and Python languages, SQL queries and MongoDB aggregations, *.csv* files, *.xls* spreadsheets, SPSS statistical analysis).

The main data processing flow diagram is given in Figure 17.

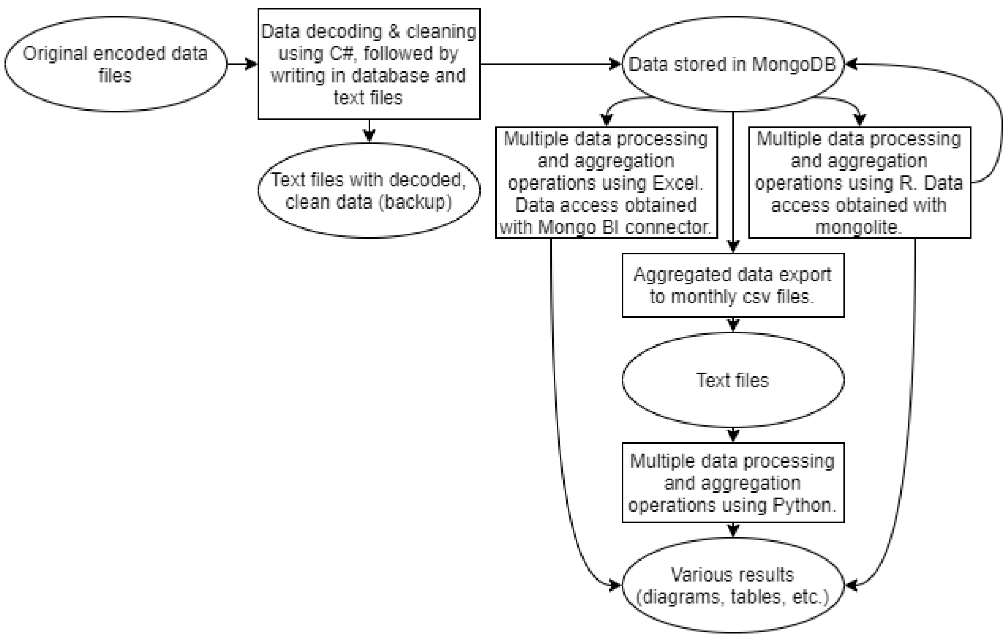

Figure 17. Data processing flow.

4. Methodology for Analyzing the Questionnaire's Data

Every data mining process requires the sequential completion of the following essential steps: problem statement, data collection, survey and preprocessing, data modeling and knowledge deployment. The survey is a commonly used method for gaining insights into data so the appropriate modeling tools can be used. Multiple strategies might be applied in order to preprocess the data, and sometimes different data sets are needed, therefore visualizations and meaningful summaries are required.

SOM are models used to lower the dimensional space in which data is represented, especially when the number of clusters is unknown. The SOM approach reorganizes data into a lower- dimensional space, by transferring similar data into corresponding areas. Usually, neurons are placed at the nodes of a two-dimension lattice and the locations of the nodes (coordinates) are indicators of the statistical characteristics contained in the input vector of that specific neuron. The purpose of the SOM is to convert the incoming signal pattern of random dimension into a one-or-two-dimensional discrete map and to adapt this transformation to a topologic order as in Figure 18.

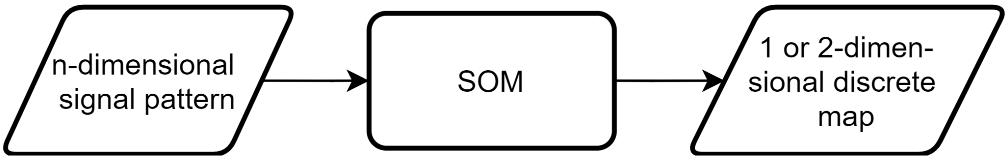

Figure 18. Transforming incoming signal with Self Organizing Maps (SOM).

Each neuron will be connected to the nodes from the input vector. The dimension of the input space is denoted by n, therefore a randomly selected vector from the input space will look as follows:

$$x = [x_1, x_{2,\ldots,} x_n]^T \tag{1}$$

The weight space of each neuron in the network will have the same dimension as the input vector.

$$w_i = \left[w_{i_1}, w_{i_{2\ldots}}, w_{in}\right]^T \tag{2}$$

where $i \in \{1, 2, \ldots., p\}$, and p is the number of neurons from the network.

To obtain the best pair of the input vector with the weight vector w_i, the products ${w_i}^T \times x$ will be calculated and the largest one will be selected, a method which is equivalent to minimizing the Euclidean distance between x and w_i. The result will identify the topologic region (neighborhood) where the excited neurons are centered. Identifying the index of the neuron that best matches the input space means the fulfillment of the following condition:

$$i(x) = \mathrm{arg}min\|x - w_i\| \tag{3}$$

The neuron identified is called the winning neuron or the winner-takes-it-all neuron for the input vector x. Provided the need of the application, the network will return either the index of the winning neuron or the weight vector that is closest in Euclidian terms to the input space. The topological neighborhood for neuron i (the winning neuron is center of the topological area as in Figure 19), and the excited neuron (we consider a specific one j), denoted $h_{j,i}$, is a function of the distance between the excited neuron and the winning neuron, denoted $k_{j,i}$, defined by two requirements: $h_{j,i}$ is symmetrical with the point $k_{j,i}$ = 0; and the amplitude of the function decreases while the distance increases.

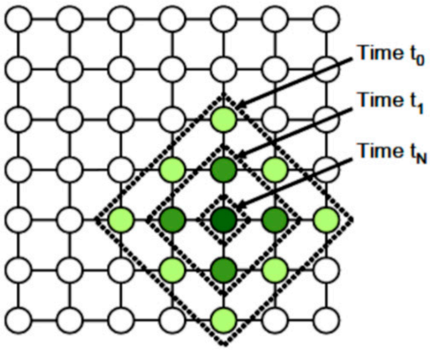

Figure 19. Neighbor selection of the winning neuron [28].

Multiple measurements methods for distance calculation might be used such as: Manhattan distance, vector product, Mahalanobis or Chebyshev distance.

One common use for the topological neighborhood function that satisfies these two requirements is the Gaussian function, where σ parameter defines the extent to which neurons participate in the learning process or the neighborhood radius at iteration i.

$$h_{j,i(x)} = \exp(-\frac{{k_{j,i}}^2}{2\sigma^2}) \tag{4}$$

The last step in the process of SOM algorithm definition is the synaptic adaptive process, therefore the vector w_j of the neuron j must change in relation to the input x. The equation for updating the weight vectors is as follows:

$$w_i(t+1) = w_i(t) + \varphi(t)h_{ij}(t)\left[x_j - w_i(t)\right] \tag{5}$$

where i is the index of Best Matching Unit (BMU), t is the learning step and φ is the learning rate parameter.

The learning rate parameter decreases gradually as time increases. The neighborhood size decreases also as the iteration number increases.

The methodology for implementing the unsupervised SOM algorithm is presented in Figure 20.

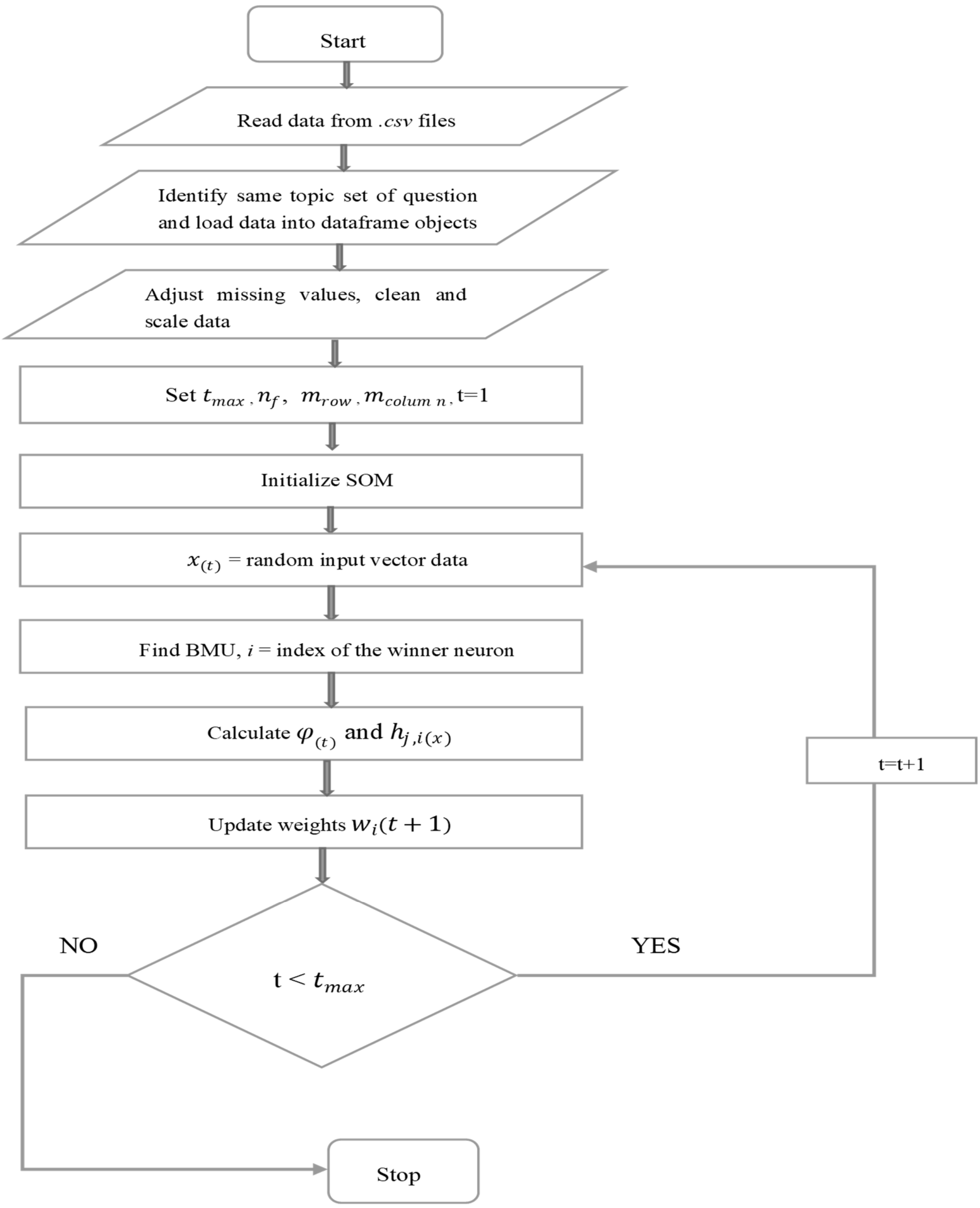

Figure 20. Methodology of unsupervised SOM algorithm on questionnaire data.

5. Results

5.1. Electricity Consumption Data

The consumption data summing up more than five hundred million rows were handled in MongoDB and analyzed with Python. In the following paragraphs, we depict the consumption data. In this sense, the total hourly consumption, for all types of consumers for each weekday, is shown in Figure 21. It is obvious that weekend profiles are slightly different than the rest of the weekdays.

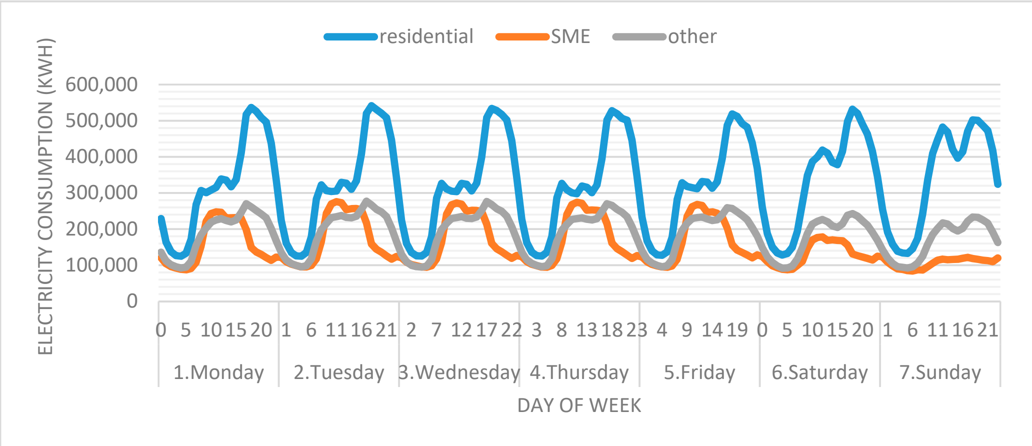

Figure 21. Total consumption (kWh) per category of customers, weekday and hour.

The total hourly consumption decomposition for residential groups and for each weekday is presented in Figure 22. In this case, load profile of groups A and C are similar. Also, B and D groups' load profiles are similar, while E is in between the two groups.

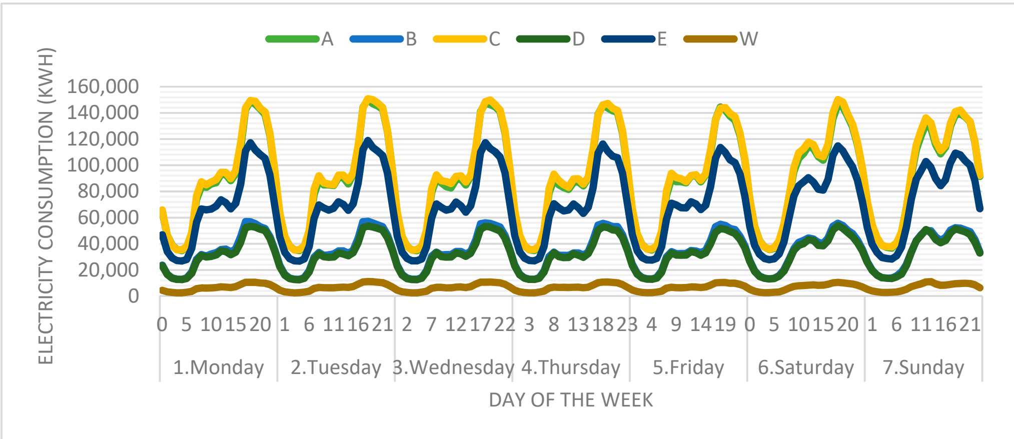

Figure 22. Total hourly consumption (kWh) of residential consumers, per tariff, for each weekday.

The proportion and the total consumption of consumers' categories is given in Figure 23. The majority (66%) consists in residential consumers. Almost half of the total consumption belongs to residential consumers.

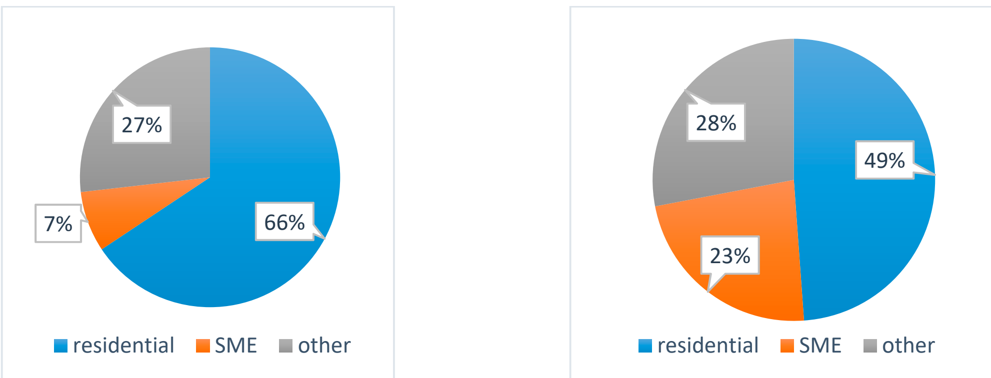

Figure 23. The percentage and total consumption (kWh) per category of customers.

The average consumption for each group of residential consumers is described in Figure 24. Except for D and E groups, the averages are similar. While, at first glance, the figure may seem to indicate significant variations between D and E categories and the rest, the difference is small enough to ignore / attribute to statistical variations (3.76% between D and W, and 4.69% between E and W).

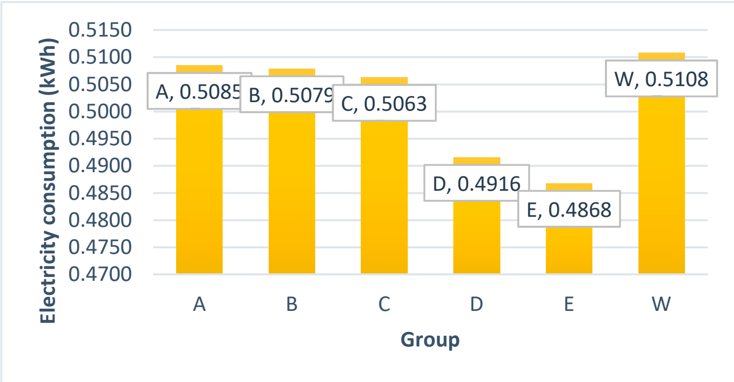

Figure 24. The average hourly consumption (kWh) of a residential customer per tariff.

The daily load profile as total consumption for the three categories is given in Figure 25. The differences are evident, and they are dependent on the activities of each consumers.

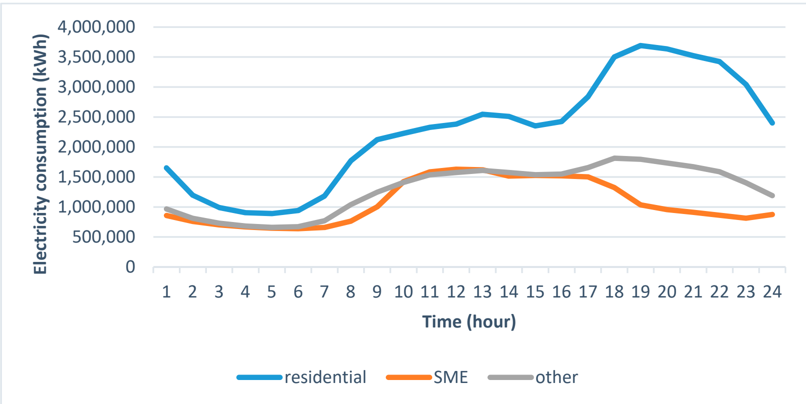

Figure 25. The daily load profile (kWh) of residential, small and medium enterprises (SMEs) and other consumers.

The daily load profile as total consumption for the six residential groups is given in Figure 26. The shapes are similar especially for A and C or B and D groups, while E profile is in between of the two groups and W profile is almost flat.

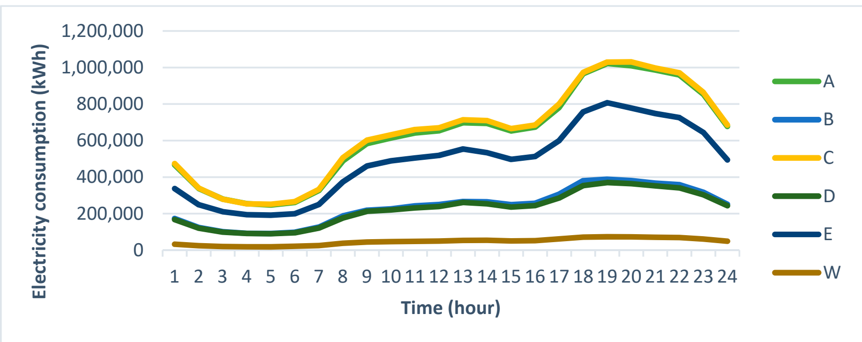

Figure 26. The average hourly consumption (kWh) of a residential customer per tariff.

The heatmap in Figure 27 shows the average hourly consumption level for each test group that has allocated a certain tariff. It reveals several aspects:

- Actual peak, off-peak and mid-peak hours do not identically correspond with ToU rates. For instance, the peak hours stretch from 17 to 21;
- D and W groups have the lowest average hourly consumption probably as a consequence of high peak rate;
- A and B groups have the highest peak consumption due to the less punitive peak rates.

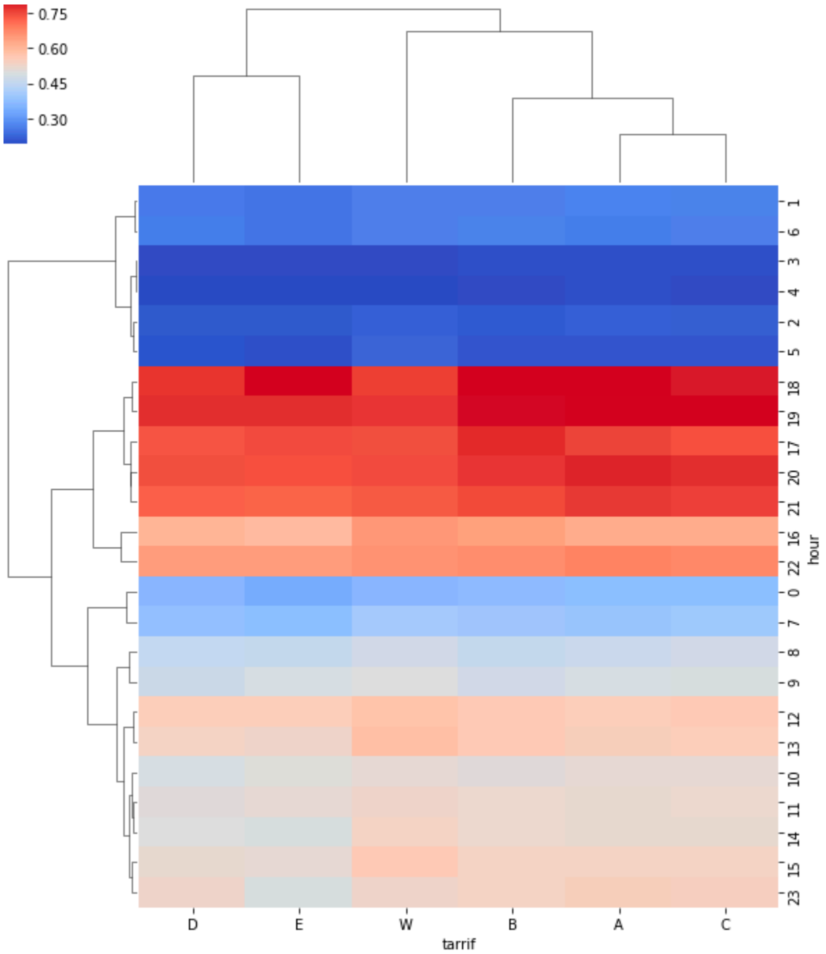

Figure 27. Heatmap test groups vs. hourly consumption.

5.2. *ToU Tariffs*

The ToU tariff are multiplied by the hourly consumption of each consumer (identified by a meter ID) for entire period from December 2009 to November 2010. The data set is processed in a dataframe (df) format in Python Pandas library.

$$\sum_{meterID} \sum_{1}^{31} \sum_{h=0}^{23} C_h \times ToU_{Xy} = pX \tag{6}$$

Five ToU tariffs $X = \{A, B, C, D, W\}$ with different rates are characterized by three levels $y = \{peak,\ off-peak, mid-peak\}$.

Intervals associated to the rates are described as follows: peak hours: 18, 19 $\rightarrow$ 2 hours; off-peak: 0–8, 20–23 $\rightarrow$ 13 hours; mid-peak: 9–17 $\rightarrow$ 9 hours.

Several *what-if* scenarios are carried out simulating that a test group of consumers X pays other tariffs $\rightarrow monthly\_pX \in \{pA,\ pB,\ pC,\ pD, pW\}$ The payment for a test group is calculated considering the five ToU tariff rates.

The optimal payment is the minim value of the payments with different ToU tariffs at the monthly level. The comparisons are monthly performed as the measures regarding house energy efficiency are gradually implemented during the trial period.

$$opt_{payment} = \min(monthly\_pX) \tag{7}$$

The difference between the payment with the initial tariff E and ToU tariffs are evaluated calculating a monthly coefficient c_m.

$$c_m = \frac{pX}{pE} \times 100\% - 100 \tag{8}$$

The reduction coefficient to improve the ToU tariffs are calculated as average of the c_m.

$$c_r = average(c_m) \tag{9}$$

The residential consumers were initially allocated the ToU tariffs forming six groups that corresponds to each tariff as in Figure 28. The purpose of applying various ToU tariffs was to test if, and in what measure, the consumers can be persuaded, via tariff, to change their consumption behavior. In addition, the electricity bill amount was observed as the payment is a significant incentive.

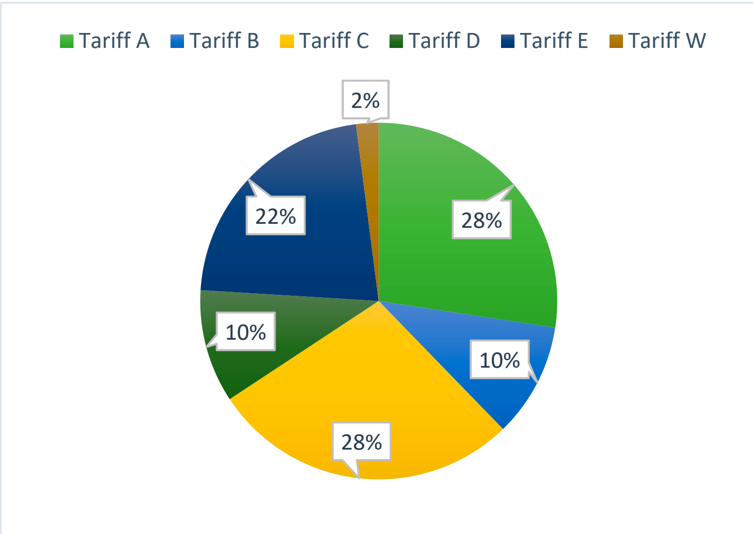

Figure 28. Initial allocation of ToU tariffs.

However, the recommended tariffs that would minimize the electricity payment considerably differ. For this analysis, we performed monthly what-if scenarios that lead to the conclusion that in most of the cases tariff A and W are recommended, since they minimize the consumers' payment, as shown in Figure 29.

Considering the frequency of recommended ToU tariffs, we conclude that only for group W, the allocated tariff minimized the payment in 50% of the time (for 6 months), while for group A, the tariff minimized the payment for 7 from the 12 months. For groups C and D, the more convenient option is tariff A or W, mainly because W has higher peak rate similar with their allocated tariffs. The advantage of tariff W is the flat lower tariff rate applied on weekend days. Thus, it advantages the consumers groups with the high consumption on the weekend. Also, A is an efficient tariff for all test groups (especially A and B), as it has the lowest peak rate.

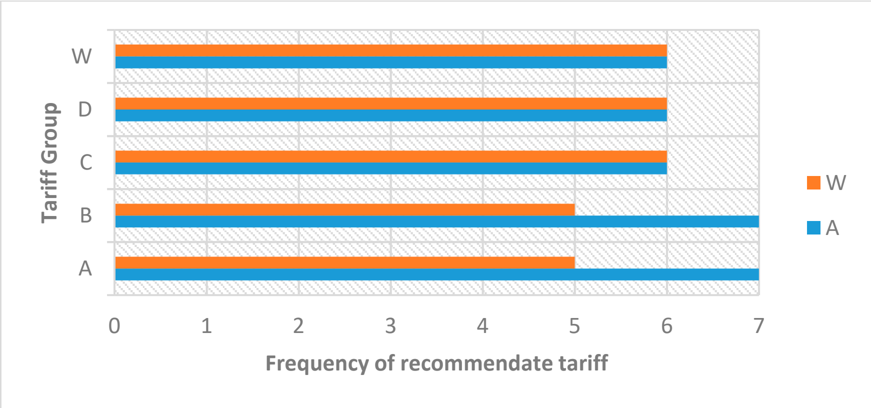

Figure 29. Frequency of recommended ToU tariffs.

Thus, we simulate that each group of consumers would pay each of the proposed tariffs, and identified the tariffs that minimize the payment, as in Figure 30.

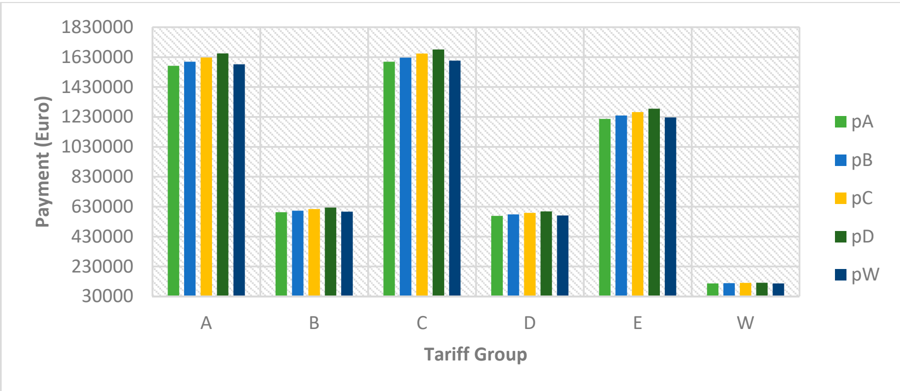

Figure 30. Payment with each ToU tariffs at test groups level.

We also simulated the payment with each of the ToU tariffs for all consumers, regardless of the test group at the monthly level. Figure 31 shows higher payment during winter months that is influenced by a higher consumption for heating. However, the lowest payment is obtained with tariff A (seven times) and W (five times).

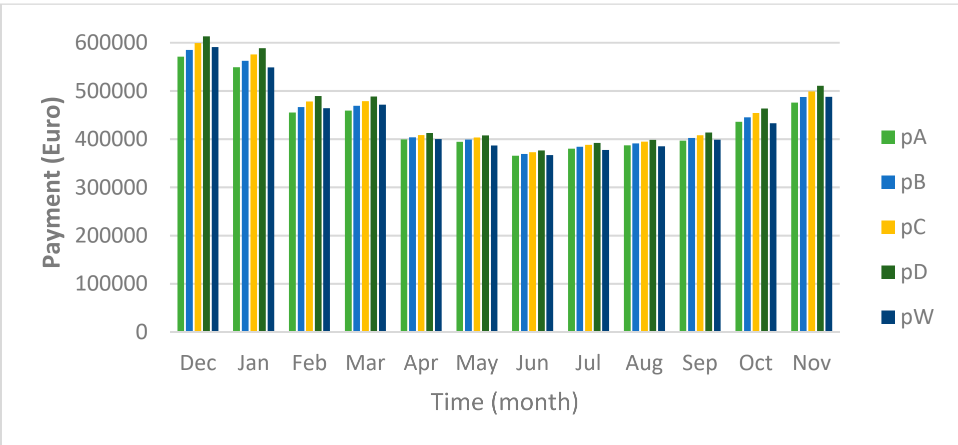

Figure 31. Payment with each ToU tariffs at monthly level.

In other words, tariffs B, C and D, with higher peak rates, could not be recommended as higher payments result with these tariffs. Still, most of the consumers are better-off with tariff A that has the lowest peak rate. Although tariff W has the highest peak rate during the working days, on weekend days the rate is flat and lower, corroborated with a high consumption, as in Figure 32.

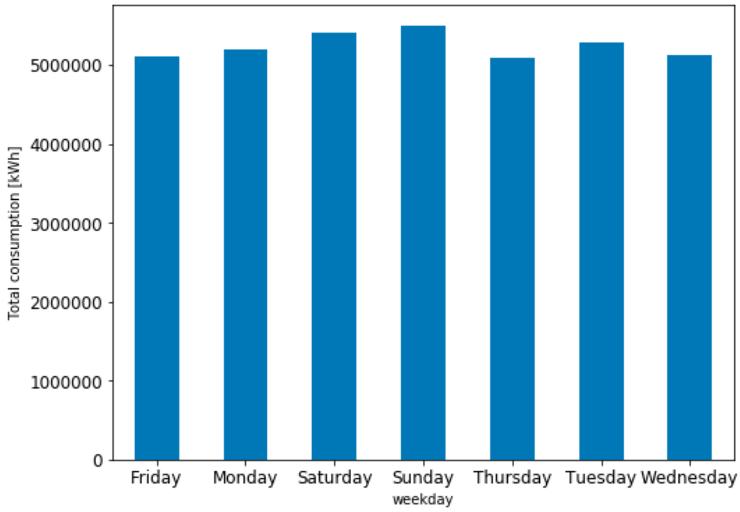

Figure 32. Electricity consumption for each weekday.

Figure 33 shows the differences between payment with tariff E and payment with the allocated ToU tariffs for each month.

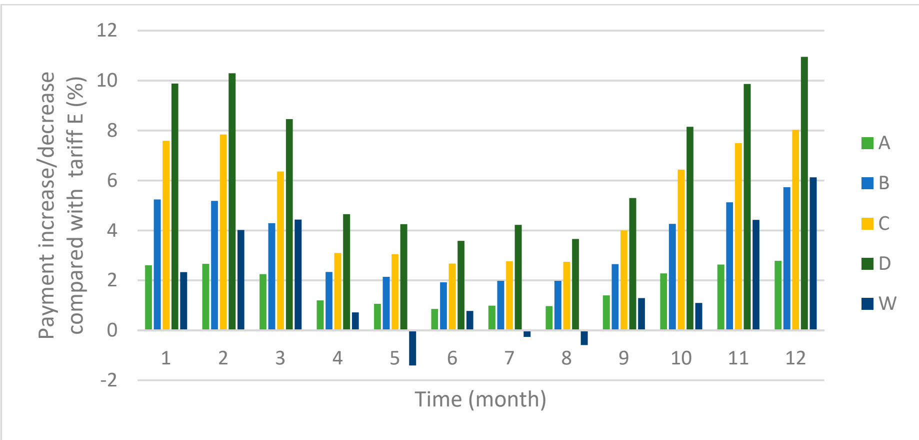

Figure 33. Monthly difference between payment with tariff E and payment with ToU tariffs.

On average, the consumers' payment was with around 19.39% higher than with tariff E. In only three months—June, August and September—tariff W proved more efficient than tariff E.

Additionally, Figure 34 shows the variation of ToU tariffs along a year, reveling that concentric circles for A–D tariffs, while W, with the butterfly shape is different crossing the A and B circles for certain months (December and March), mainly due to the weekend low rate.

Calculating the mean of these differences of payment between tariff E and ToU tariffs, we identified the reduction that will improve the ToU tariffs as in Table 1.

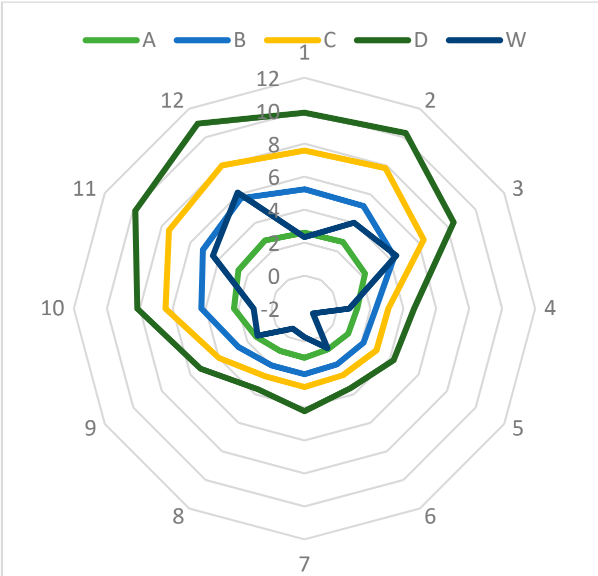

Figure 34. Monthly radar difference between payment with tariff E and payment with allocated ToU tariffs.

Table 1. Tariff reduction.

Tariff	%Reduction
A	1.80
B	3.57
C	5.17
D	6.94
W	1.91

While Figure 26 seems to suggest that there are differences in the consumption profiles of the various customers categories, per ToU tariff, this is only due to the variated size of the categories, and not to their consumption patterns. Figure 35, which shows the average consumption for each customers category, per tariff, indicates only slight variations between the categories.

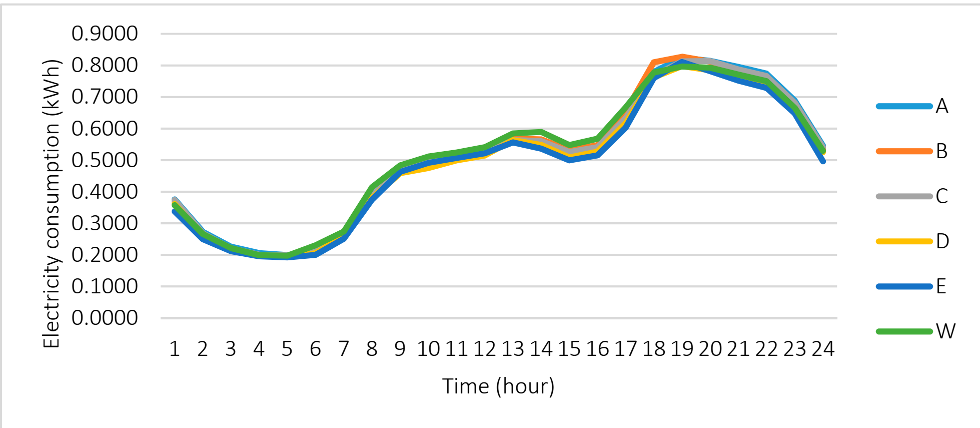

Figure 35. The average hourly consumption (kWh) of residential customers per tariff.

Regarding the data the Figure 35 is based on, the conclusion of our analysis is that the proposed method for changing the customers' consumption behaviors via differential tariffs could not provide clear evidence. The consumers followed their inherent consumption patterns without any regard for the penalties imposed, in various degrees, by most of the ToU tariffs, on the consumption at peak hours. This gives a large variation in the total electricity to be produced by generators and transported/distributed by the grid operators, per hour, with a minimum in consumption, at 4'o clock (in the night), of 890,219 kWh and a maximum, at 18 (in the evening), of 3,691,177 kWh. The mitigation of this large variation was exactly the reason the ToU tariffs were proposed in the first place.

Following this conclusion, we attempted to identify if the bulk of the peak hour consumption can be attributed to any particular group, not regarding the ToU tariffs, as it was obvious that there are no real differences caused by the tariffs. We found that the same consumers, which are ranked highest by the total consumption are also ranking highest by the consumption at peak hours (17–22).

First, we ordered the consumers descending, by their total consumptions, obtaining the results as in Table 2.

Table 2. Segments of consumers with high share in the total consumption.

% of Total Consumption	Segment of Consumers on the Ordered Table	% of Total Residential Customers (4225)
10% of total consumption	first 187 consumers	4.43%
25% of total consumption	first 558 consumers	13.21%
50% of total consumption	first 1336 consumers	31.62%

Next, on the same table, without any reordering, we attempted to identify the customers which are contributing more to the consumption at peak hours and found that they are almost the same, as given by Table 3.

Table 3. Segments of consumers with high share in the peak consumption.

% of 17–22 Consumption	Segment of Consumers on the Same Ordered Table	% of Total Residential Customers (4225)
10% of 17–22 consumption	first 196 consumers	4.64%
25% of 17–22 consumption	first 567 consumers	13.42%
50% of 17–22 consumption	first 1340 consumers	31.72%

Our conclusion was that it would be possible to notably alter the total consumption hourly pattern, by changing the consumption behavior of a reduced set of customers (less than one third of them, for a radical change).

We further analyzed several *what-if* scenarios, based on the idea of changing the consumption behavior of various sized selected subsets. The proposed change was to move about 50% of the consumption of the selected consumers from the peak hours (17–22) to the off-peak interval (1–6, during the night). While direct modification of consumption behaviors may not be possible, for a small enough number of customers technical approaches may be found (e.g., small electricity accumulators).

The results of the what-if scenarios can be seen in Figure 36.

The consequences of attaining the proposed scenarios are given in Table 4:

Table 4. Results of the proposed altered pattern for the higher-ranking consumers.

Indicator	Actual Total Consumption (kWh)	Results of the Proposed Altered Pattern for the Higher-ranking Consumers		
		196 (4.64%)	567 (13.42%)	1340 (31.72%)
Minimal consumption per hour	890,219	1,067,640	1,333,912	1,652,316
Maximal consumption per hour	3,691,178	3,509,498	3,233,016	2,836,683

Figure 36. The results of the proposed what-if scenarios.

As per the values given in Table 4, if such results are achievable, the best-case consequence would be a 23.15% reduction on the maximal consumption could be sustained.

5.3. Questionnaire Insights

Clusters are one of the most efficient ways to represent features of data, and the SOM algorithm is especially relevant for analyzing survey data, because of its visualization properties. When data is unlabeled, the algorithm is efficient by indicating the number of classes, but if data is labeled, the algorithm may be used for dimensionality reduction.

The algorithm creates one or more prototype-vectors that are relevant for the input data set, and it preserves the topology of the data by projecting the set of the prototype vectors from the dimensional space onto a low-dimensional grid. Pre- and post-data surveys contain opened and closed questions related to consumer profile, consumption trends, but especially related to the attitude towards consumption and the considerations related to reducing electricity consumption.

The questionnaire responses were loaded into pandas' Data Frames objects (questions were represented as columns and each respondent *id* defined the index for each row) and various functions were applied in order to determine the data types, the number of missing responses or statistical insights. Data preprocessing implied replacing inaccurate data with significant values, in accordance with the question type, as described in the data processing steps in Figure 37. Also, some redundant questions were removed.

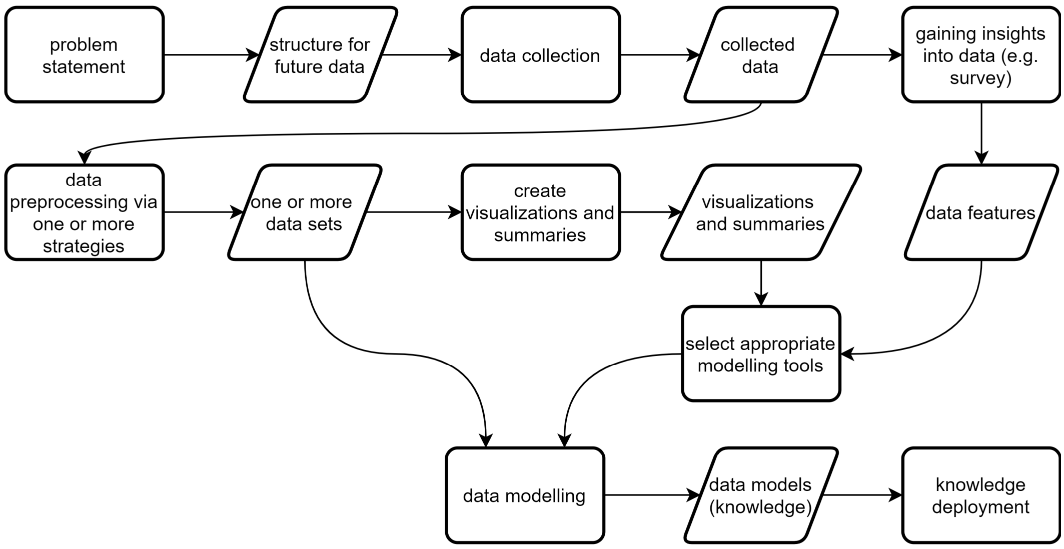

Figure 37. Steps for data processing.

For the vast majority of the questions, scaling was not necessary because of the question type—binary or categorical—and also because of the meaningful codification of answers where this was relevant. For some questions, the standard scaler was used, and for a few features with large magnitudes, a min-max scaler was applied.

The network was trained for 10,000 epochs with a learning rate of 0.01 and with a sigma of 1. One feature is selected (question: "*I [we] can reduce my [our] electricity bill by changing the way the people I live with and I [we] use electricity*)", and the distribution of nodes is presented in Figure 38.

The map in Figure 38 reflects the network nodes for a subset of the post-questionnaire set that includes answers related to the attitude in relation to the reduction of energy consumption. Answers related to the personal assessment of the knowledge of reducing consumption were considered in the input data. The SOM visualization helps to identify the classes of consumers, using as input space the attitude type questions towards a certain situation, such as: the society/individual must or should reduce the consumption of electricity, or the motivation behind consumption reduction: environmental problems or personal financial reasons.

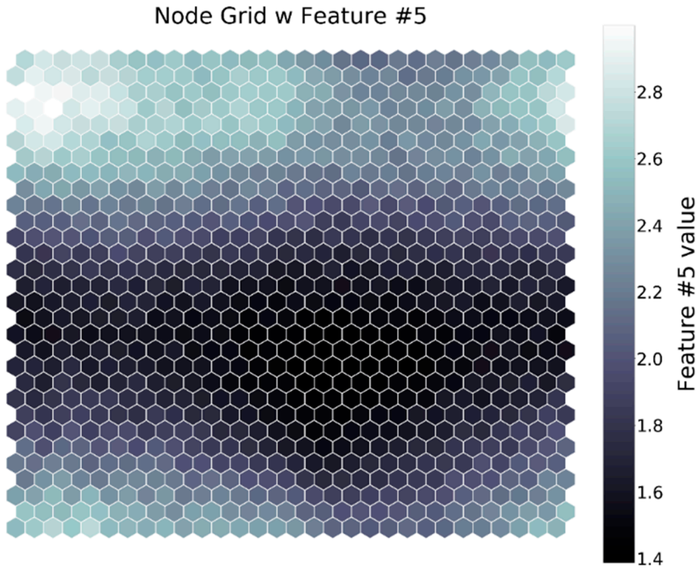

Figure 38. Map of nodes colored in accordance with the fifth feature on a 30x30 map.

A Unified-Index Matrix represents a special graph type that reflects the distance between the nodes in the grid. A large distance is represented by a dark area, while the lighter colored areas mean a smaller distance between nodes-edges between similar data groups. For an input space of dimension 11, the features were the answers to questions about the household income and the answers to the question *"I/we have made changes to the way I/we live in order to reduce the amount of electricity I/we use"*. After generating a 30 x 30 SOM, in which each vector represents one or more items, the U-Matrix was constructed as in Figure 39 by computing the sum of the Euclidian distances for each neighboring cell and calculating the average. If the result is small, then the items more likely belong to the same class.

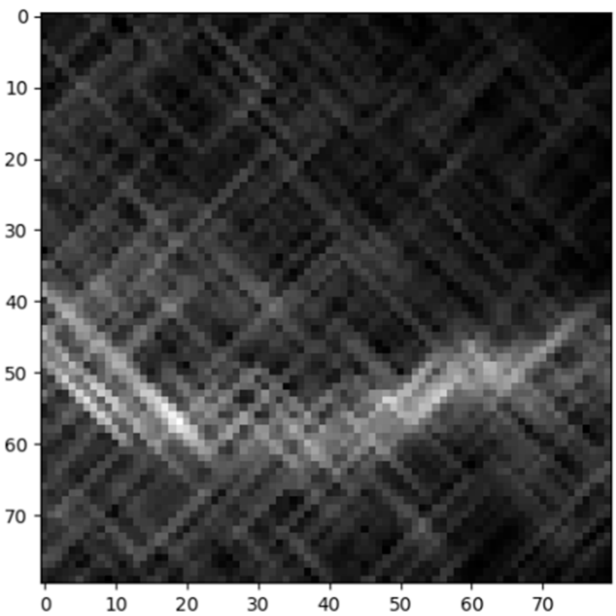

Figure 39. U-Matrix.

By looking at the U-matrix, the lines suggest that there are four areas of similar consumers. Dimensionality reduction is graphed in Figure 40.

We can deduct that items from the blue area are very different than the items in brown area and green-blue and brown-yellow areas are somewhat similar. Also, we can observe one SOM limitation, namely that categorical answers are not handled well, because the algorithm assumes that the variables are continuous. In the same time, inconsistent solutions were identified while running the analysis multiple times, because initial positions of neurons differ. The cluster number can only be determined after the algorithm consistency was established.

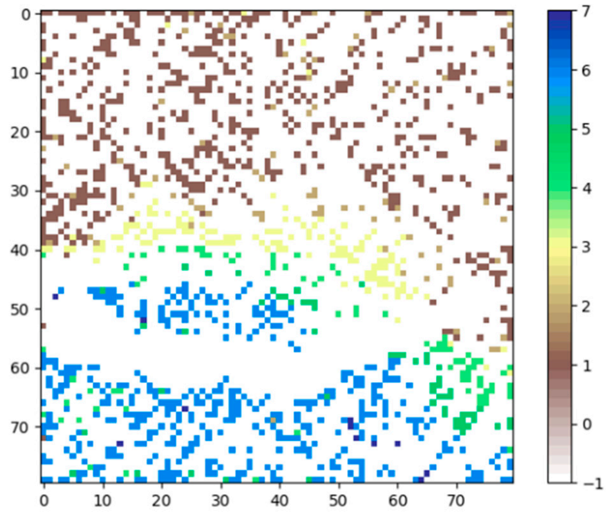

Figure 40. Dimensionality Reduction.

Another limitation is that the number of iterations is difficult to be determined, but according to [28] the map will converge, after an adequate number of iterations. We also set up a group of questions, named set of questions 1, related to the same topic: "*perception of the usefulness of the instruments received at the beginning of the trial (monitors, stickers, magnets, etc.)*".

For all these questions the answers are on a scale from 1 to 6, where, to questions such as: "*how useful were the stickers or magnets*", 1 as answer means totally useless, and 6 means very useful. The missing answers were filled with 0, and also scaling was applied to improve accuracy, because some questions only had the scale of the answers from 1 to 5.

This set of questions also contains questions linked to the electricity monitor, from the evaluation of the time of understanding of the device's operating mode (1:very easy, and 6:very difficult) to the bill evaluation in terms of electricity consumption. It can easily be observed that questions with a similar response (black squares as in Figure 41) have been arranged by the neural network very close to each other. Black squares represent the answer to the question codified by 5—which means strongly agreed or very satisfied with the outcomes. The default color for missing values (zeros) was red colored, but it was removed for a better and more understandable representation. Incidentally, other answers are represented close to each other (1-green, 2-cyan, 4-white,3-blue, 6-yellow). The white ones are in immediate closure to the black squares, which means the network identified correctly groups of respondents that agree or are satisfied with the electricity monitor and may consider that over the trial the amount of electricity was reduced. Other markers—blue, green and cyan are also close to each other, but more scattered over the map, which means that some of the respondents have strong beliefs in report to some situations or questions, but they disagree on other statements.

For another set of questions with a topic related to the person's own measures taken to reduce consumption, the magnitude of these measures, the degree of modification of the consumption mode (day/night or hourly intervals), the grouping of answers is represented in Figure 42.

We can see that the respondents who are represented by the yellow marker (6: strongly agree) are very well delimited by those who gave answers from 1 to 4. It can be deduced that those who have adopted their own measures (minor or major measures) have also observed a decrease in energy consumption and reported a general change in consumption mode.

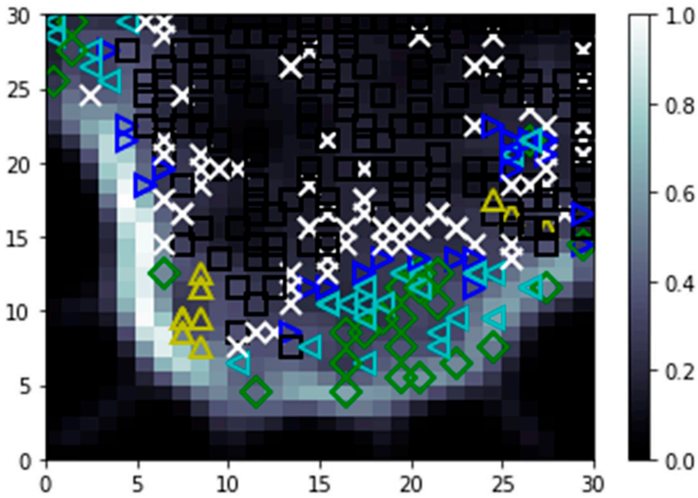

Figure 41. Winner nodes representation for set of questions 1.

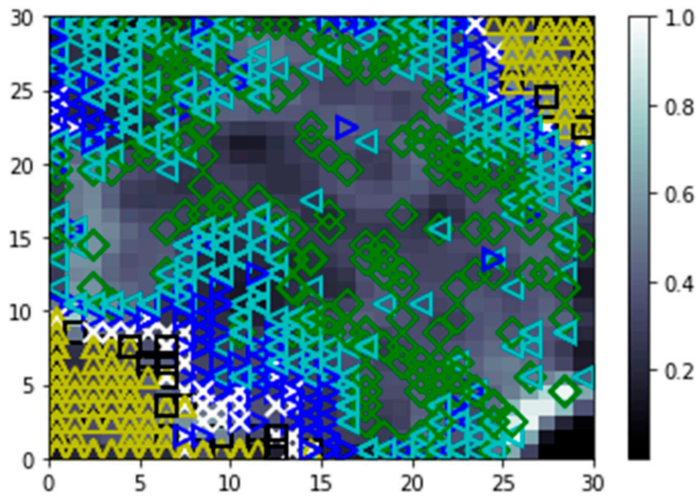

Figure 42. Winner nodes representation for set of questions 2.

6. Discussion

In this section, we analyzed a couple of references and the current study (as in Table 5) from different point of view, comparing the sample size in terms of number of households and the main findings of the researches considering their target that was briefly described in Section 2.2.

Compared to the above mentioned research papers, the current study analyzes complex and high-dimensional datasets, such as: a large electricity consumption dataset in correlation with ToU tariffs implemented during the trial period after the installation of SM, pre- and post-trial voluminous questionnaires that revel the consumers' pattern and the behavioral trend.

Table 5. Comparison among several references and the current study.

No.	Reference	Summary	Country	Sample Size (Households)	Main Findings
1	[13]	Energy consumption feedback - installing real-time display of electricity consumption	Australia and other countries	Variability of sample size is analysed: larger sample sizes are correlated with lower conservation effects	A small reduction in the consumption level, around 3–5%, together with a reduction of the consumption peak, but it did not lead to lower carbon or greenhouse gas emissions
2	[14]	Installing different types of devices to monitor real-time energy consumption in residential houses	United States	151	A reduction of 12% in mean electricity consumption
3	[18]	Use more performant feedback technologies, providing real-time information that is accessible through mobile devices	Switzerland	N.A.	5.7% average reduction in consumption, especially at peak hours
4	[23]	Effects of real-time feedback provided by the consumers, together with a set of information regarding energy saving measures	Austria	1500	An average of approx. 4.5% economy in electricity consumption
5	[24]	Effects of smart display for households	United States	432	Short-term electricity consumption dropped by 9%
6	[25]	Effect of introducing ToU tariffs for residential consumers	Italy	1446	Peak in the morning was diminished and the bills were reduced by 2.21%, the average electricity consumption increased by 13.69%
7	[26]	Identify measures and causes that can influence the reduction of residential consumption	Sweden	2000	Consumers that had web available consumption display achieved energy savings around 15%
8	Current study	Analyse complex datasets: electricity consumption, tariffs and questionnaires	Ireland	4224	23% of the peak consumption can be reduced by shifting it from peak to off-peak hours

7. Conclusions

In this paper, input data analysis on several data sets starting with consumption data, tariffs and surveys were performed.

On one hand, the consumption data recorded by more than 4000 SM indicate the load profile of the residential consumers for weekdays and weekend days. The most recommended ToU tariffs proved to be tariff A, with the mildest peak rate, and tariff W that has a very convenient rate during the weekend.

On the other hand, the effects of the proposed tariffs in correlation with the consumption data were investigated, leading to the improvement of the ToU tariffs so as to minimize the electricity expenses. Also, the consumption at peak was investigated, revealing that segments of consumers majorly influence the peak consumption. For instance, we discovered that 50% of the peak consumption belong to about 1300 consumers, which is less than one third of the total consumers. Hence, DSM should be directed to these segments that have the highest impact of the load curve.

We clustered the consumers with SOM based on the answered from the pre- and post- trial surveys. A major advantage of the SOM algorithm is that it is an intuitive method of segmenting the profiles of the questionnaire respondents. A disadvantage for the study of the questionnaires with SOM is the coding of the open questions, because the training stage implies the existence of numerical data. Standardization is not always necessary, but it improves the numerical accuracy of codified responses. Kohonen's algorithm is simple, yet powerful in processing and analyzing survey data, as it by provides important data information by placing significant attributes in an input data set in a small grid. The algorithm is trained in a unsupervised manner on a large set of input data, and if these contain groups, the data vectors matched by these groups are mapped by SOM, so that the distribution of the vectors is an approximation of the distribution of the original data set. Map visualization significantly improves data understanding.

Author Contributions: All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This paper presents the scientific results of the project "Intelligent system for trading on wholesale electricity market" (SMARTRADE), co-financed by the European Regional Development Fund (ERDF), through the Competitiveness Operational Programme (COP) 2014–2020, priority axis 1 – Research, technological development and innovation (RD&I) to support economic competitiveness and business development, Action 1.1.4-Attracting high-level personnel from abroad in order to enhance the RD capacity, contract ID P_37_418, no. 62/05.09.2016, beneficiary: The Bucharest University of Economic Studies.

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

List of variables

data	Dataframe objects containing set of questions with answers, columns are the questions and respondents' ids are the indexes of the rows. Questions sets might be identified by applying topic modelling technique.
n_f	Number of features
m_{row}	Number of rows on map
m_{column}	Number of columns on map
t	Iteration index
t_{max}	Number of epochs
$x_{(t)}$	Random selected input
$i_{(x)}$	Index of the BMU

$\varphi_{(0)}$	Starting learning rate
$h_{j,i(x)}$	Neighborhood function that determines the distance between excited neuron *j* and winning neuron *i*
$\varphi_{(t)}$	Learning rate at step *t*. Learning rate decrease as iteration increases.
$w_i(t)$	Weight of node *i* at step *t*
C_h	Hourly consumption
X	Tariff type
y	Tariff level
pX	Payment with a ToU tariff (A, B, C, D or W)
$monthly\_pX$	Monthly payment with a ToU tariff
$opt_{payment}$	Optimal payment
c_m	Monthly coefficient
c_r	Reduction coefficient
T_{Xy}	A certain tariff type level

List of acronyms

BMU	Best Matching Unit
DSM	Demand Side Management
ICT	Information Communication Technology
MR	MapReduce
NoSQL	Not only Structured Query Language
SM	Smart Meters
SME	Small and Medium Enterprises
SOM	Self-Organizing Maps
ToU	Time-of-Use

References

1. Boyes, H.; Hallaq, B.; Cunningham, J.; Watson, T. The industrial internet of things (IIoT): An analysis framework. *Comput. Ind.* **2018**. [CrossRef]
2. O'Neill, J. *Demand Response: Electricity Market Benefits and Energy Efficiency Coordination*; Energy Policies, Politics and Prices; Nova Science Publishers, Incorporated: New York, UK, 2014; ISBN 9781629480732.
3. Comission for Energy Regulation—CER Electricity Smart Metering Customer Behaviour Trials (CBT) Findings Report. Available online: https://www.cru.ie/wp-content/uploads/2011/07/cer11080ai.pdf (accessed on 14 January 2020).
4. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**. [CrossRef]
5. Aurélien, G. *Hands-on Machine Learning with Scikit-Learn & TensorFlow*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017; ISBN 9781491962299.
6. AtHebb, D.O. The Organization of Behavior; A Neuropsychological Theory. *Am. J. Psychol.* **1950**. [CrossRef]
7. Kohonen, T.; Somervuo, P. Self-organizing maps of symbol strings. *Neurocomputing* **1998**. [CrossRef]
8. Becker, S. Unsupervised learning procedures for neural networks. *Int. J. Neural Syst.* **1991**. [CrossRef]
9. Kohonen, T.; Somervuo, P. Self-organizing maps of symbol strings with application to speech recognition. In Proceedings of the WSOM, Espoo, Finland, 4–6 June 1997.
10. Tsai, C.W.; Lai, C.F.; Chao, H.C.; Vasilakos, A.V. Big data analytics: A survey. *J. Big Data* **2015**. [CrossRef]
11. He, Y.; Tan, H.; Luo, W.; Mao, H.; Ma, D.; Feng, S.; Fan, J. MR-DBSCAN: An efficient parallel density-based clustering algorithm using MapReduce. In Proceedings of the International Conference on Parallel and Distributed Systems—ICPADS, Tainan, Taiwan, 7–9 December 2011.
12. Han, D.; Agrawal, A.; Liao, W.K.; Choudhary, A. A novel scalable DBSCAN algorithm with spark. In Proceedings of the 2016 IEEE 30th International Parallel and Distributed Processing Symposium, IPDPS 2016, Chicago, IL, USA, 23–27 May 2016.
13. McKerracher, C.; Torriti, J. Energy consumption feedback in perspective: Integrating Australian data to meta-analyses on in-home displays. *Energy Effic.* **2013**. [CrossRef]

14. Alahmad, M.A.; Wheeler, P.G.; Schwer, A.; Eiden, J.; Brumbaugh, A. A comparative study of three feedback devices for residential real-time energy monitoring. *IEEE Trans. Ind. Electron.* **2012**. [CrossRef]
15. Frederiks, E.R.; Stenner, K.; Hobman, E.V. Household energy use: Applying behavioural economics to understand consumer decision-making and behaviour. *Renew. Sustain. Energy Rev.* **2015**, *41*, 1385–1394. [CrossRef]
16. Geelen, D.; Reinders, A.; Keyson, D. Empowering the end-user in smart grids: Recommendations for the design of products and services. *Energy Policy* **2013**. [CrossRef]
17. Goulden, M.; Bedwell, B.; Rennick-Egglestone, S.; Rodden, T.; Spence, A. Smart grids, smart users? the role of the user in demand side management. *Energy Res. Soc. Sci.* **2014**. [CrossRef]
18. Houde, S.; Todd, A.; Sudarshan, A.; Flora, J.A.; Armel, K.C. Real-time feedback and electricity consumption: A field experiment assessing the potential for savings and persistence. *Energy J.* **2013**. [CrossRef]
19. Kendel, A.; Lazaric, N. The diffusion of smart meters in France: A discussion of the empirical evidence and the implications for smart cities. *J. Strateg. Manag.* **2015**. [CrossRef]
20. Khan, A.R.; Mahmood, A.; Safdar, A.; Khan, Z.A.; Khan, N.A. Load forecasting, dynamic pricing and DSM in smart grid: A review. *Renew. Sustain. Energy Rev.* **2016**, *54*, 1311–1322. [CrossRef]
21. Maréchal, K.; Holzemer, L. Getting a (sustainable) grip on energy consumption: The importance of household dynamics and "habitual practices". *Energy Res. Soc. Sci.* **2015**. [CrossRef]
22. Mizobuchi, K.; Takeuchi, K. The influences of financial and non-financial factors on energy-saving behaviour: A field experiment in Japan. *Energy Policy* **2013**. [CrossRef]
23. Schleich, J.; Klobasa, M.; Gölz, S.; Brunner, M. Effects of feedback on residential electricity demand-findings from a field trial in Austria. *Energy Policy* **2013**. [CrossRef]
24. Schultz, P.W.; Estrada, M.; Schmitt, J.; Sokoloski, R.; Silva-Send, N. Using in-home displays to provide smart meter feedback about household electricity consumption: A randomized control trial comparing kilowatts, cost, and social norms. *Energy* **2015**. [CrossRef]
25. Torriti, J. Price-based demand side management: Assessing the impacts of time-of-use tariffs on residential electricity demand and peak shifting in Northern Italy. *Energy* **2012**. [CrossRef]
26. Vassileva, I.; Odlare, M.; Wallin, F.; Dahlquist, E. The impact of consumers' feedback preferences on domestic electricity consumption. *Appl. Energy* **2012**. [CrossRef]
27. Berkhin, P. A survey of clustering data mining techniques. In *Grouping Multidimensional Data: Recent Advances in Clustering*; Springer: Berlin, Heidelberg, Germany, 2006; ISBN 354028348X.
28. Abhinav Ralhan Self Organizing Maps. Available online: https://towardsdatascience.com/self-organizing-maps-ff5853a118d4 (accessed on 3 February 2020).