Grade Analysis for Households Segmentation Based on Energy Usage Patterns

Tomasz Ząbkowski
Department of Informatics
Warsaw University of Life Sciences
Warsaw, Poland
tomasz_zabkowski@sggw.pl

Krzysztof Gajowniczek
Department of Informatics
Warsaw University of Life Sciences
Warsaw, Poland
krzysztof_gajowniczek@sggw.pl

***Abstract*—The Grade Correspondence Analysis (GCA) with posterior clustering and visualization is introduced and applied to individual households' electricity usage data. The main task of this analysis is to identify a way of representing the variability of a households behavior and to develop an efficient way of clustering the households into a few, usable and homogenous groups. The regularity in terms of the electricity usage is useful information for organizations to allow accurate demand planning with the aim of improving the overall efficiency of the network. The approach is tested using data from 46 households located in Austin, Texas, USA and monitored for 14 months at a sampling interval of 1 hour.**

Keywords—grade correspondence analysis; households segmentation; electricity usage patterns; smart meter data

I. INTRODUCTION

This paper proposes GCA segmentation approach applied to electricity usage patterns of individual customers. The knowledge of the loads variability at different groups of electricity customers is very important to assure proper operation management of power distribution networks. The changes taking place in the electricity market require effective methods to provide the end users profiles and their demand for the electricity which is the basis for formulating pricing strategies, constructing tariffs and undertaking actions to improve the efficiency and reliability of distribution networks. Load patterns categorization for tariff purposes is usually performed on aggregate residential load data, or on individual non-residential load data. Residential consumers are generally not considered as individual entities due to the fact that consumption patterns of individual residential customers vary which is the function of the number of inhabitants, their activity, age and lifestyle [1]. Various techniques for customer classification are presented and discussed in the literature, with the focus on highlighting the behavior of electricity customers [2], [3], [4]. The core stage of the categorization process is the use of appropriate clustering techniques to perform load pattern grouping [1], [5], [6]. Comparisons among the techniques have been carried out with regard to various clustering validity indicators [1], [6], [7].

In particular, the proposed paper addresses some aspects of transversal grouping which is aimed at grouping the load patterns of residential customers whose data have been gathered in similar conditions and covering the same time span. The grouping of the customers will be based taking into account historical electricity consumption data and, additionally, the household's behavioral factors that impact the energy usage of individual appliances observed at the household level. We believe such enhanced set of explanatory variables can significantly improve the segmentation analysis. In particular, the scope of the paper is twofold:

1) Application of grade cluster analysis to identify the consumption patterns of the customers and grouping together customers exhibiting similar patterns; Additional information on operating selected home appliances is considered to deliver value added in terms of more homogeneous segments created.

2) Visualization of the segments on the over-representation maps to support the interpretation of the results.

The proposed research fits into the attempt focused on leveraging smart meter data to support energy efficiency on the individual user level. This poses novel research challenges in monitoring usage, gathering data, controlling devices, communicating with the energy grid and inferring from data [8], [9]. In the attempt of reducing electricity consumption in buildings, identification of homogenous patterns of energy consumption at different customer groups is a key to improve efficiency of available energy usage. In this context, there are attempts which are focused on appliance recognition taking into account the use of home infrastructure such as non-intrusive appliance load monitoring approach (NIALM) [10], [11]. This is to identify trends in the use of domestic appliances from household electricity consumption measurements since customer profiling is methodically sound and offers a variety of potentials for application within the energy industry [12], [13], [14].

In the following sections we characterize the data and introduce the basics of grade analysis. Subsequently, we describe the technical and methodological realization as well as the evaluation using real data. The final section provides summary and an outlook on further application scenarios.

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

II. SMART METER DATA USED

Smart metering systems are the components influencing and creating environmental sustainability by managing energy at homes. They are supposed to be important factor in reducing overall energy consumption and increasing energy awareness of the end users through being better informed about the their consumption patterns. In order to be perceived as intelligent, smart metering systems need to have the following features: (1) automatic processing, transfer, management and utilization of metering data; (2) automatic management of meters; (3) two way data communication protocol with meters; (4) be able to provide meaningful and timely consumption information to the utilities and end-users; (5) should support services for energy, water or heating savings.

For the purpose of this research we have used WikiEnergy dataset [15] by Pecan Street Inc. which is a large database of consumer energy information. This database is highly granular, including the usage measurements collected from up to 24 circuits within the home. The investigated households were located in Austin, Texas, USA. From these readings, we have extracted 14 months of data from 46 households at a granularity of 1 hour, covering the same time window, which was from March 2013 until April 2014.

Recently, with advances in communication infrastructure for remote and automated data reading, there has been increasing interest in analysis of residential power loads. However, patterns of electricity use at a system demand level and at an individual level are very different. For instance, Fig. 1 shows the pattern of electricity use for a single random dwelling extracted from the WikiEnergy data. The profile shows two peaks in the morning at 6 am and 8 am and the afternoon peak that is smaller than the peak in the morning between 4 pm and 6 pm. In contrast, Fig. 2 reflects a distinctly different pattern of electricity use for a group of 46 households on the same day of the year. The figure shows a smooth profile shape with relatively little electricity consumption in the early afternoon, a clearly defined peak in the morning and a slightly smaller defined peak in the evening.

As presented in Fig. 3 there are quite huge differences observed among the households in the analyzed dataset. The average load ranges from 0.31 kWh to 3.64 kWh with standard deviation of 0.366 kWh and 2.362 kWh, respectively.

In addition, for the purpose of the analysis, specific behavioral data were extracted, indicating the on-off statuses of the devices. The reference data were collected for: washing machine (WM), dish washer (DW), tumble dryer (TD), microwave (MV) and small kitchen devices including kettle (KE). This was the base to calculate the average length of the time window between subsequent turn on of each of the devices.

III. GRADE DATA ANALYSIS

Grade data analysis is efficient technique that works on variables measured on any measurement scales, including categorical. It bases on dissimilarity measures such as concentration curves and precisely defined measure of monotonic dependence. The method's framework is constituted of grade transformation proposed by [16] and developed by [17], [18]. The underlying idea is to transform any distribution of two variables into a convenient form of the so called grade distribution. This transformation leaves unchanged the order of variables, ranks and the values of monotone dependence measures like Spearman's ρ^* and Kendall's τ. In case of empirical data this approach consists of analyzing the two-way table with rows/columns, which is preceded by proper recoding of variable values.

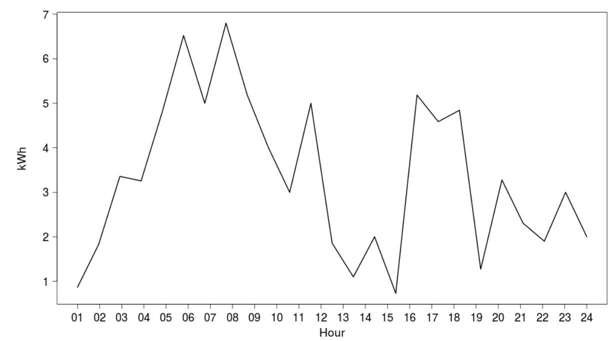

Fig. 1 Daily electricity demand load profile across a 24 hr period on 21st July 2013 for an individual dwelling.

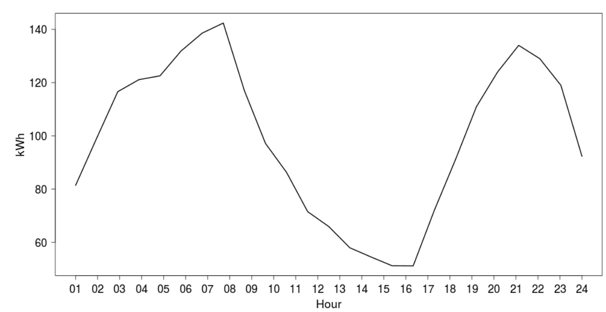

Fig. 2 Daily electricity demand load profile across a 24 hr period on 21st July 2013 aggregated for 46 households.

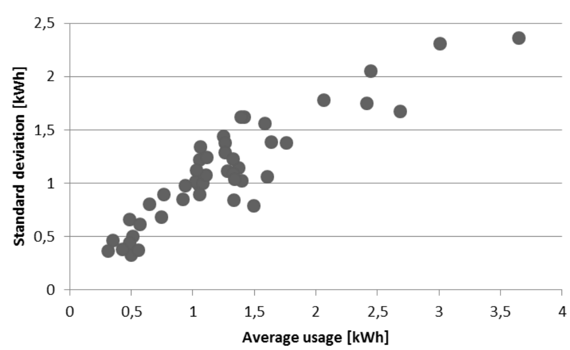

Fig. 3 The relation between average usage (in kWh) and standard deviation at each of 46 househols.

The main tool of the grade methods is Grade Correspondence Analysis (GCA), which benefits from classical correspondence analysis, however it is going significantly beyond it, by means of grade transformation. GCA is ordering the columns and row of the table in such a way that neighboring rows are more similar than those further apart, and at the same time, neighboring columns are also more similar than those further apart. After the optimal ordering is found it is possible to aggregate neighboring rows and neighboring columns, and therefore, to build clusters with similar distributions. The Spearman ρ* was originally defined for continuous distributions, but it may be defined also as Pearson's correlation applied to distribution after the grade transformation. The grade distribution may be defined for discrete distribution too, and it is possible to calculate Spearman ρ* for probability table P with m rows and *k* columns, where p_{is} is the frequency (treated as probability) of *i-th* row in *s-th* column:

$$\rho^*(\mathrm{P}) = 3\sum_{i=1}^{m}\sum_{s=0}^{k}\left(p_{is}\left(2S_{row(\mathrm{i})}-1\right)\left(2S_{\mathrm{col(s)}}-1\right)\right) \tag{1}$$

where

$$S_{row(\mathrm{i})} = \left(\sum_{j=1}^{i-1} p_{j+}\right)+\frac{1}{2}p_{i+}, S_{\mathrm{col(s)}} = \left(\sum_{t=1}^{s-1} p_{+t}\right)+\frac{1}{2}p_{+s} \tag{2}$$

and p_{j+} and p_{+t} are marginal sums defined as: $p_{j+} = \sum_{s=1}^{k} p_{js}$, $p_{+t} = \sum_{t=1}^{m} p_{ts}$.

GCA tends to maximize ρ* by ordering row and columns according to their grade regression value, which is the center of the gravity for each row or each column. The grade regression for the rows is defined as:

$$regr_{row(i)} = \frac{\sum_{s=1}^{k} p_{is}S_{\mathrm{col(s)}}}{p_{i+}} \tag{3}$$

and for the columns:

$$regr_{\mathrm{col(s)}} = \frac{\sum_{i=1}^{m} p_{is}S_{\mathrm{row(i)}}}{p_{+s}}. \tag{4}$$

The algorithm calculates the grade regression for columns and sorts the columns by its values what results in increase of the regression for columns, but at the same time the regression for rows changes. If the regression for rows is sorted then regression for columns changes. As proved in [17] each sorting of the grade regression increases the value of Spearman ρ*. The number of possible states (combination of permutations of rows and columns) is finite and it is equal to k!m!. Each time the value of Spearman ρ* increases, and the last ordering produces the largest ρ*, called local maximum of Spearman ρ*. The output from GCA depends on the initial permutation of rows and columns, and if it is ordered in reversed way with respect to initial permutation, it is possible to achieve symmetrically reversed local maximum, which is a pair of ρ*: original and its reversal.

In consecutive steps, GCA primarily permutes randomly rows and columns and reorders them to achieve a local maximum. In practice, for the bigger $m \times k$ tables, the search over the all possible combinations of rows and columns is a long-lasting and computationally demanding process. Therefore, in order to find global maximum of ρ*, Monte Carlo simulations are used. For this purpose, starting from randomly ordered rows and columns, the algorithm is iteratively searching for such a representation where ρ* reaches local maximum. The highest value of ρ* is chosen from the set of local maxima and it is assumed to be equal to the global maximum, what is usually obtained after 100 iterations of the algorithm. It is important to underline that calculation of grade regression requires non-zero sum of every row and column in a table, so this requirement applies also to the GCA. More detailed description on grade transformation can be found in [18], [19].

When it comes to the grade cluster analysis (GCCA), it's framework is based on optimal permutations provided by the GCA. The assumption is that numbers of clusters are given, rows and/or columns of the data table (variables, say X and Y) are optimally aggregated. The respective aggregated probabilities in this table arose from sums of component probabilities in initial, optimally ordered table, and number of rows in the aggregated table equals the desired number of clusters. In this case, optimal clustering means that ρ* (X, Y) is maximal in the set of these aggregations of rows and/or columns, which are adjacent in optimal permutations. The rows and columns may be aggregated either separately – maximizing ρ* for aggregated X and non-aggregated Y or for non-aggregated X and aggregated Y, or simultaneously. Details concerning the maximization procedure can be found in [20].

Finally, grade analysis technique is aided by visualizations using over-representation maps. The maps serve as a very convenient tool for plotting data structures. Every cell in the data table is represented by the respective rectangle in [0,1] x [0,1] and it is marked by various shades of grey, what corresponds to the level of the randomized grade density. The value range of the grade density is divided into several intervals and each color represents a particular interval. The black corresponds to the highest values while the white to the lowest. As grade density measures deviation from independence of variables X and Y, so dark colors indicate overrepresentation while the light ones show underrepresentation.

IV. Clustering Experiments

The starting point for the analysis was to prepare the matrix with electricity usage per each hour and for each household. The sample structure of the dataset is presented in Table 1.

TABLE 1. THE SAMPLE MATRIX WITH AVERAGE HISTORICAL ELECTRICITY USAGE (IN KWH) PER EACH HOUR (T_1-T_24) AND FOR EACH HOUSEHOLD.

Household	t_1	t_2	t_2	…	t_24
1	0.67	0.57	0.49	…	0.85
2	0.90	0.77	0.89	…	1.27
…	…	…	…	…	…
46	0.51	0.43	0.38	…	0.63

The second step was to prepare the matrix with average length of the time window between subsequent turn on of each of the devices. The sample structure of the dataset prepared for the analysis is presented in Table 2.

TABLE 2. THE SAMPLE MATRIX WITH THE TIME ELAPSED (IN HOURS) BETWEEN SUBSEQUENT APPLIANCES TURN ON EVENTS, FOR EACH HOUSEHOLD.

Household	WM	DW	TD	MV	KE
1	97.8	139.0	52.6	11.4	2196.0
2	10.1	10.7	14.3	25.0	9.8
…	…	…	…	…	…
46	23.4	14.5	22.1	4.7	1.4

Let's take for instance the household no. 1. The average time between the use of each of the appliances is as follows: nearly 98 hours for washing machine, 139 hours for dishwasher, nearly 53 hours for tumble dryer 11 hours for microwave and about 91 days for the kitchen devices. This indicates that the home is rather rarely occupied.

The combined data structures presented in Table 1 and Table 2 has been analyzed in GradeStat tool [21] which was developed in Institute of Computer Science Polish Academy of Science.

The first step was to calculate over-representation ratios for each field (cell) of the table. A given $m \times k$ data matrix with non-negative values can be visualized using over-representation map in the same way as a contingency table [20]. Instead of frequency n_{ij} the value of *j-th* column for *i-th* row is used. Next, it is compared in a contingency table with the corresponding neutral or fair representation $n_{i\bullet} \times n_{\bullet j} \big/ \sum\sum n_{ij}$ where $n_{i\bullet} \big/ \sum_j n_{ij}$, $n_{\bullet j} \big/ \sum_i n_{ij}$. The ratio of the first and second expression is called the over-representation ratio. An over-representation surface over a unit square is divided into $m \times k$ rectangles situated in m rows and k columns, and the area of rectangle placed in row i and column j being equal to fair representation of normalized n_{ij} . Having the over-representation ratios the over-representation map for the initial raw data can be constructed. The color of each field in the map depends on the comparison of the two values: (1) the real value of measure connected to the considered field and corresponding to population element; (2) the expected value of the measure. Fig. 4 presents initial over-representation map for the analyzed data. The cells' colors in the map are grouped into three classes:

(1) *gray* – the measure for the element is **neutral** (ranging between the 0.99-1.01) what means that the real value of the measure is equal to its expected value;

(2) *black* or *dark gray* – the measure for the element is **over-represented** (between 1.01 and 1.5 for weak over-representation and more than 1.5 for strong) what means that the real value of the measure is greater than the expected one;

(3) *light gray* or *white* the measure for the element is **under-represented** (between 0.66 and 0.99 for weak under-representation and less than 0.66 for strong under-representation) what means that the real value of measure is less than the expected one.

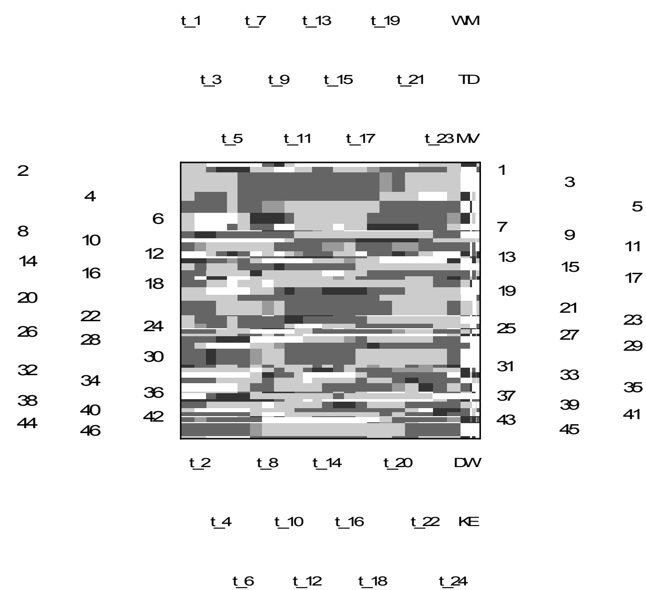

Fig. 4 The initial over-representation map.

The following step was to apply the grade analysis to measure the dissimilarity between two data distributions (households and usage variables dimensions) in order to reveal the structural trends in data. The grade analysis was done based on Spearman's ρ*, used as the total diversity index. The value of ρ* strongly depends on the mutual order of the map's rows and columns. To calculate ρ*, the concentration indexes of differentiation between the distributions are used. The basic procedure of GCA is executed through permuting the rows and columns of a table in order to maximize the value of ρ*. After each sorting the ρ* value increases and the map becomes more similar to the ideal one. That means that the darkest fields are placed in the upper-left and lower-right map corners while the rest of the fields is assigned according to the following property: the farther from the diagonal towards the two other map corners (the lower-left and upper-right ones) the lighter gray (or white) color the fields have.

The result of the GCA procedure is presented in Fig. 5. Additionally, cluster analysis was performed through the aggregation of some rows representing unique households. The optimal number of four clusters was obtained when the changes of the subsequent ρ* values appeared to be negligible as referenced in [8], [18]. In Fig. 6 the chart with the ρ* values as a function of the number of clusters is presented. The points on the OX axis correspond to the cluster numbers. The OY axis is denoted by the values of ρ*.

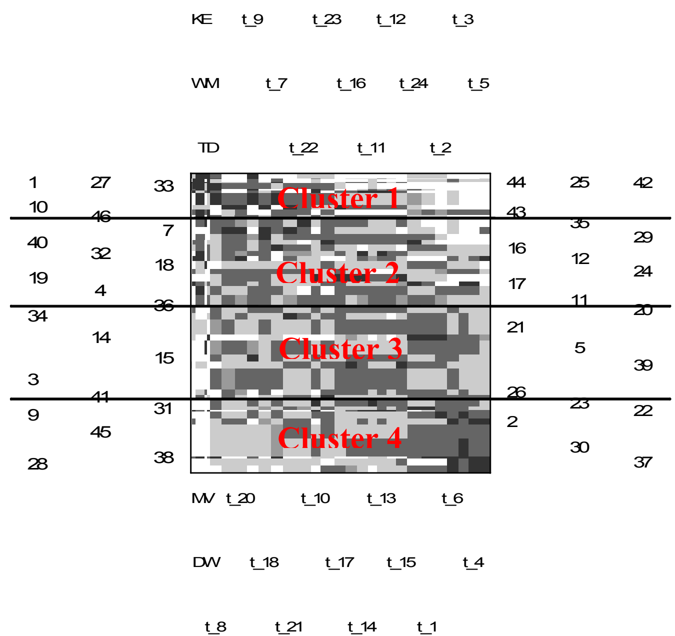

Fig. 5 The final over-representation map with four clusters.

The resulting order presents the structure of underlying trends in data. Four clusters show typical behavior of the households in terms of the electricity usage and the way of operating appliances. Cluster 1 consists of 12 households (1, 6, 8, 10, 13, 25, 27, 33, 42, 43, 44, 46) and it can be characterized by strong over-representation in terms of the frequent usage of selected home appliances taking place in the morning (t_7, t_8, t_9) and in the evening (t_18, t_20). Cluster 2 consists of 14 households (4, 7, 11, 12, 16, 17, 18, 19, 24, 29, 32, 35, 36, 40) and in comparison to Cluster 1 it can be characterized by the less frequent usage of selected home appliances. Similarly to Cluster 1, the electricity usage takes place over the whole day, except the night hours (t_1 – t_6).

The Cluster 3 is made of 10 households (3, 5, 14, 15, 20, 21, 26, 34, 39, 41) and it can be characterized by strong under-representation in terms of the usage of selected home appliances. In general, the electricity usage characteristic of the segment is spread over the whole day, including night hours.

Finally, the Cluster 4 contains 10 households (2, 9, 22, 23, 28, 30, 31, 37, 38, 45). The characteristics of the behavior suggest very occasional use of selected home appliances and very intensive (over-represented) electricity consumption during the night hours (t_1 – t_6).

Additionally, k-means clustering was performed to have some reference with respect to GCA. However, the results revealed that the method tends to create only two clusters as confirmed by Calinski-Harabasz index [22] and Rousseeuw's Silhouette internal cluster quality index [23].

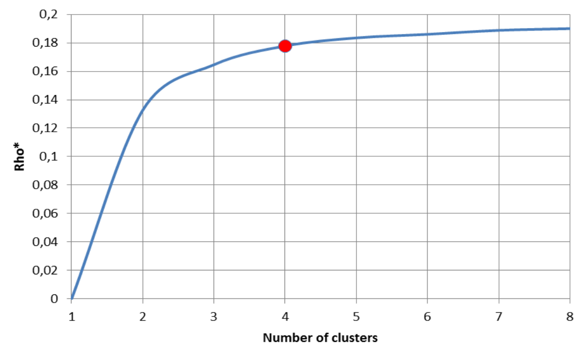

Fig. 6 The ρ* values for different number of clusters.

TABLE 3. STRUCTURAL AND BUILDING SPECIFIC CHARACTERISTICS OF THE CLUSTERS.

Cluster	Building types	Average construction year	Average footage	Photovoltaics systems *
1	Majority of Apartments (>50%)	2004	1640	No
2	Majority of Single Family homes (>50%)	2007	2112	Yes
3	Only Single Family homes	1998	2480	No
4	Single Family homes and large Apartments	2007	2058	Yes

*Yes/No depending on whether the majority of buildings (>50%) is benefiting or not from the program.

The results of the GCA clustering were validated against the structural and building specific data. Those were: building type (Apartment, Single-Family Home, Town Home), construction year, total square footage, and incentive program due to installed photovoltaics systems. The results are presented in Table 3. The analysis confirmed that derived clusters are characterized by different structural and building specific data. Cluster 1 consists of relatively new apartments with average footage of 1640, usually not benefiting from photovoltaics systems.

Cluster 2 consists of relatively new single family houses with average footage of 2112 and benefiting from photovoltaics systems. Cluster 3 gathers older single family houses with the biggest average footage of approx. 2480 and not profiting from the incentive program due to installed photovoltaics systems.

Cluster 4 comprises of new single family houses and large apartments with average footage of 2058 and benefiting from photovoltaics systems.

The proposed GCA method applied for the clustering enables identification of the customers' consumption patterns and then grouping together those exhibiting similar patterns. The knowledge of the loads variability at different customer groups is very important to assure proper operation management of power distribution networks. Additionally, the results may have the implications for the demand response and efficiency programs.

V. CONCLUSIONS

In this paper, we proposed a way to generalize the observed usage patterns for the group of 46 households. The goal of the analysis was to identify a way of representing the variability of households behavior and to develop an efficient way of clustering the households into a few, usable and homogenous groups.

Grade exploration allowed for quickly grasping general trends in data, and then to cluster the households taking into account historical usage supported by the information on operating selected home appliances. The data obtained by grade analysis might be the basis for the decision support to consider both: the price elasticity tariffs and accurate forecasting of the electricity load, for the identified homogenous groups of customers.

Since the results are promising and visually appealing we aim to explore algorithmic approaches for mining households usage patterns which can have applications in predicting future behavior or developing unique, individualized energy management strategies for selected groups of the customers. The literature studies revealed that there is a gap in existing studies of association or sequential rule learning in the context of energy consumption, despite extensive works on web usage and recommendation systems, for instance. Therefore, capturing households specific patterns and usage preferences using a rule learning approach could produce, in our opinion, clear rules intelligible to the general audience. Those can be further utilized to develop clusters of homogenous electricity users.

ACKNOWLEDGMENT

The study was cofounded by the National Science Centre, Poland Grant No. 2016/21/N/ST8/02435 and WULS-SGGW International Research Scholarship Fund No. INT-282/2016.

REFERENCES

[1] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping", *Energy*, 42(1), 2012, 68–80.

[2] K. Gajowniczek, and T. Ząbkowski, "Data Mining Techniques for Detecting Household Characteristics Based on Smart Meter Data", *Energies*, vol. 8, 2015, 7407–7427.

[3] D. Jang, J. Eom, M. J. Park, and J. J. Rho, "Variability of electricity load patterns and its effect on demand response: A critical peak pricing experiment on Korean commercial and industrial customers", *Energy Policy*, vol. 88, 2016, 11–26.

[4] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data", *IEEE Transactions on Smart Grid*, vol. 7(1), 2016, 136–144.

[5] G. Chicco, R. Napoli, F. Piglione, M. Scutariu, P. Postolache, and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 19(2), 1232–1239, May 2004.

[6] I. Dent, T. Craig, U. Aickelin, and T. Rodden, "An approach for assessing clustering of households by electricity usage", *Proc. 12th Annu. Workshop Comput. Intell. (UKCI)*, 2012, Available at SSRN: https://ssrn.com/abstract=2828465

[7] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer", *IEEE Trans. Power Syst.*, vol. 18(1), 381–387, Feb. 2003.

[8] J. Carroll, S. Lyons, and E. Denny, "Reducing household electricity demand through smart metering: The role of improved information about energy saving", *Energy Economics*, vol. 45, 2014, 234–243.

[9] M. Anda and J. Temmen, "Smart metering for residential energy efficiency: The use of community based social marketing for behavioural change and smart grid introduction", *Renewable Energy*, vol. 67, 2014, 119–127.

[10] G. M. Tina and V.A. Amenta, "Consumption awareness for energy savings: NIALM algorithm efficiency evaluation", *5th International Renewable Energy Congress (IREC)*, IEEE, 2014.

[11] I. Dent, "Deriving knowledge of household behaviour from domestic electricity usage metering", *PhD Thesis*, University of Nottingham, 2015.

[12] F. Rodrigues, J. Duarte, V. Figueiredo, Z. Vale, and M. Cordeiro, "A comparative analysis of clustering algorithms applied to load profiling", *Proc. of MLDM 2003*, Leipzig, Germany, 73–85, 2003.

[13] S. Ramos, and Z. Vale, "Data mining techniques to support the classification of MV electricity customers", *IEEE Power and Energy Society General Meeting – Conversion and delivery of electrical energy in the 21st century*, 2008, 1–7.

[14] T. Zabkowski, K. Gajowniczek, and R. Szupiluk, "Grade analysis for energy usage patterns segmentation based on smart meter data", *IEEE 2nd International Conference on Cybernetics (CYBCONF),* June 2015.

[15] Pecanstreet, Dataport. 2014. Available from https://dataport.pecanstreet.org/

[16] W. Szczesny, "On the performance of a discriminant function", *Journal of Classification* , vol. 8, 1991, 201–215.

[17] A. Ciok, T. Kowalczyk, E. Pleszczyńska, and W. Szczesny, "Algorithms of grade correspondence-cluster analysis", *Archiwum Informatyki Teoretycznej i Stosowanej*, vol. 7, 1995, 5–22.

[18] T. Kowalczyk, E. Pleszczyńska, and F. Ruland (Eds.) „Grade Models and Methods of Data Analysis. With applications for the Analysis of Data Population", *Studies in Fuzziness and Soft Computing*, vol. 151, Springer Verlag, Berlin - Heidelberg - New York, 2004.

[19] W. Szczesny, "Grade correspondence analysis applied to contingency tables and questionnaire data", *Intelligent Data Analysis*, vol. 6(1), 2002, 17–51.

[20] A. Ciok, T. Kowalczyk, and E. Pleszczyńska, "How a New Statistical Infrastructure Induced a New Computing Trend in Data Analysis", Polkowski L., Skowron A. (Eds.) *Rough Sets and Current Trends in Computing*, *Lecture Notes in Artificial Intelligence*, vol. 1424, 1998,75–82.

[21] GradeStat, Program for Grade Data Analysis, www.gradestat.ipipan.waw.pl

[22] T. Caliński, and J. Harabasz, "A dendrite method for cluster analysis", *Communications in Statistics-theory and Methods*, vol. 3(1), 1974, 1–27.

[23] P.J. Rousseeuw, and "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, vol. 20, 1987, 53–65.