Electricity demand profile prediction based on household characteristics

Joaquim L. Viegas, Susana M. Vieira, João M. C. Sousa
IDMEC, LAETA, Instituto Superior Técnico, Universidade de Lisboa, Portugal
joaquim.viegas@tecnico.ulisboa.pt

R. Melício
IDMEC, LAETA, Instituto Superior Técnico, Universidade de Lisboa, and Dep. de Física, Escola de Ciências e Tecnologia, Universidade de Évora, Portugal

V. M. F. Mendes
Instituto Superior de Engenharia de Lisboa, and Dep. de Física, Escola de Ciências e Tecnologia, Universidade de Évora, Portugal

***Abstract*—This work proposes a methodology for predicting the typical daily load profile of electricity usage based on static data obtained from surveys. The methodology intends to: (1) determine consumer segments based on the metering data using the k-means clustering algorithm, (2) correlate survey data to the segments, and (3) develop statistical and machine learning classification models to predict the demand profile of the consumers. The developed classification models contribute to make the study and planning of demand side management programs easier, provide means for studying the impact of alternative tariff setting methods and generate useful knowledge for policy makers.**

***Index Terms*—Data mining; Machine learning; Smart meter data; Household energy consumption; Segmentation.**

I. Introduction

Big changes have been happening in the utility industry and energy markets. The liberalization, growing competition between utilities, technological advancements and policy towards a sustainable use of resources are forcing utilities to seek innovation and new market related insights. Following the liberalization of energy markets, the number of utilities has been growing as has the level of competition. Utilities are becoming very invested in the research and application of new technologies, wishing to achieve higher efficiencies, lower losses, attract new customers and maintain the ones already contracted.

Technological advancements in the fields of metering, communications and computation are enabling utilities to monitor and save huge amounts of data related to their operations. The deployment of smart meters has been happening in a number of countries enabling the logging of daily costumers' consumption.

The load data of costumers has the potential to give insights of great importance for utilities. Understanding the shapes of the load curve of customers can enable the understanding of costumer habits, assist in the creation of demand side management (DSM) programs, improve load forecasting, better the efficacy of marketing campaigns and develop alternative tariff setting methods.

Due to the high number of customers and desired high sampling frequencies in smart metering, huge amounts of data are stored and its processing grows in complexity. Computational techniques in the fields of statistical and machine learning are starting to be extensively used in order to extract knowledge from the data generated by the grid.

The work proposes a methodology for: (1) the segmentation of residential electricity consumers, (2) knowledge discovery from the correlation of consumer segments and survey data and ultimately (3) the prediction of electricity demand profiles based on household characteristics.

The main aim of this methodology is to enable utilities to benefit from the knowledge related to consumer segments and their dynamics while the penetration of smart meters is still small and consumption data is only present for some costumers. The main contribution of the work is the modelling of the methodology proposed for the prediction of consumer segments based on easily obtained data.

This paper is organized as follows: in section II a small literature review of electricity consumer segmentation and of the correlation of household characteristics and energy consumption is presented; section III presents the methodology used, consisting of the profile extraction, segmentation, knowledge extraction and prediction models; in section IV the experimental results are presented; section V presents the conclusions.

II. Related Work

A. Electricity Consumer Segmentation

The segmentation of electricity consumers and load clustering has been the focus of a considerable amount of research in the past years. The usual stated applications range from the design and simulation of DSM strategies [1], [2], load forecasting [3], [4], tariff setting [5]–[7], marketing and bad data detection. The clustering methods found to be used are mostly the K-means algorithm [2], [8]–[11]. Self-organizing maps (SOM), which is also used for feature

This work was supported by FCT, through IDMEC, under LAETA, project UID/EMS/50022/2013 and SusCity (MITP-TB/CS/0026/2013). The work of J. L. Viegas was supported by the PhD in Industry Scholarship SFRH/BDE/95414/2013 from FCT and Novabase. S. M. Vieira acknowledges support by Program Investigador FCT (IF/00833/2014) from FCT, co-funded by the European Social Fund (ESF) through the Operational Program Human Potential (POPH).

978-1-4673-6692-2/15/$31.00 ©2015 IEEE

extraction [8]–[13], hierarchical clustering methods [3], [9], [14] and fuzzy clustering [13] have shown promise in the field. Data preparation is of high importance in these applications, dictating what information is desired to be extracted from the clustering and the ability of the used methods to achieve good results. Normalization, parametric modelling [4], temperature based normalization [9], [15] and wavelet transformation [3] have been found to be used in the literature.

B. *Analysis of Household Characteristics and Energy Consumption*

The use of static data related to household characteristics (e.g. income, inhabitants, education) and appliance use in relation to static or dynamic energy consumption data is being studied in order to find the main drivers of residential energy consumption. In [16]–[18] factor analysis and linear regression are used to find the main determinants of energy consumption in residential settings, such as weather data, household characteristics and demographics. In [19] demographic data and psychological and belief related data is studied in comparison to energy consumption. [20], [21] presents studies on the prediction of household information based on smart meter data. In [22], [23] consumptions profiles obtained via clustering are correlated to household characteristics in a similar fashion to what is done in this work. The main innovation of this work in comparison to the existing literature is the focus on the prediction of electricity consumption profiles based on household characteristics and not only on the analysis of the relationship between them.

III. Methodology

The methodology used in this work is shown in figure 1. The metering data is pre-processed, aggregated and filtered according to context in order to obtain the load profiles. Following this the load profiles are automatically segmented. On the other side the survey data is pre-processed and its relationship with the load profile segments is analysed using a regression model (knowledge extraction). The final step is the creation of prediction models in order to predict the segments of the consumers based on their household characteristics.

A. *Profile Extraction*

The profile extraction consists on the pre-processing, aggregation and context filtering of the metering data. The pre-processing consists on removing outliers and processing any missing values. The context filtering consists on selecting the data of a specific season, consumer group or type of day. The aggregation consists on generating a 24 hour profile from the context filtered data for each consumer (e.g. absolute mean, normalized mean and variance).

B. *Segmentation*

The K-means algorithm [24] is used in this methodology due to it being simple, efficient and scalable and proven to be adequate in this type of application. The biggest problem with this algorithm is the need to determine the number of clusters and their initial centres.

In order to automatically select the best number of clusters three cluster validity indices (CVI) are computed for multiple cluster configurations and a majority vote is used to select the best number of clusters. The CVIs represent a ratio between the cohesion of the clusters and the separation between them. The Dunn index [25] estimates the cohesion by nearest neighbour distance and the separation by the maximum cluster diameter, Davis-Bouldin index [26] estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids and the Silhouette index [27] is a summation-type index in which the cohesion is measured based on the distance between all points in the same cluster and the separation based on the nearest neighbour distance. [28]

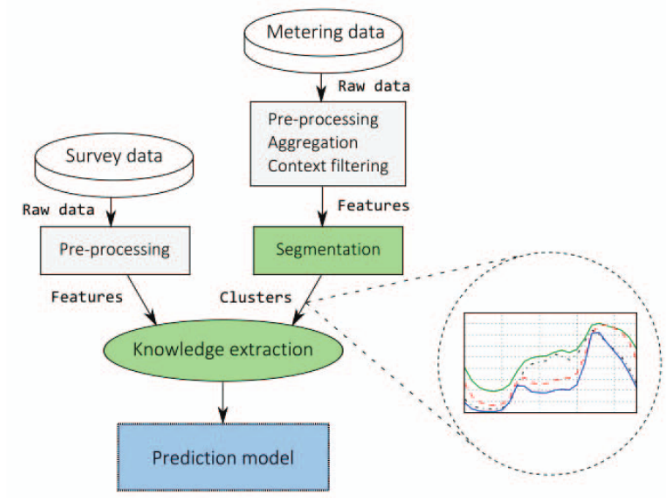


Figure 1 – Proposed methodology.

C. *Knowledge Extraction*

Based on the work of Rhodes et al. [23], the relationship between the demand profiles and survey variables is obtained using probit regression to determine if there are any significant correlations between the survey data and the probability of a consumer being in a certain cluster. The explanatory variables represent the survey data (e.g. income, number of adults, education and number of TVs) and the dependent variable is the consumer segment. The probit model is a binary classification model where the dependent variable can only take on a value of 0 or 1. In the case of the classification for more than two clusters $(n-1)$ models can be used where n is the number of clusters.

D. *Prediction Models*

In order to conclude on the ability to predict the segment of the consumption profile of a household three different models were used: probit regression, support vector machines and gradient boosted trees.

1) *Support Vector Machines Classification*

Support Vector Machines (SVMs) [29] are a popular machine learning method for classification. Given non separable training vectors in two classes Support Vector Classification (SVC) finds the hyper plane that maximizes the *margin* between the training points of classes 0 and 1, allowing some points to be inside the *margin*. The classifier finds linear boundaries in the input feature space or can make use of the kernel trick in order to work in a transformed non-linear feature space.

2) Gradient Boosted Trees

Boosting consists of combining the outputs of many "weak" classifiers to produce a powerful "committee". In gradient tree boosting the ensemble of decision trees is incrementally built by training each new model instance emphasizing the training instances that previous models misclassified.

IV. Experimental Results

This section presents the dataset used to test the presented methodology.

A. Dataset

The presented methodology was tested using the data from 4232 Irish households over a period of 1.5 years consisting of electricity consumption at 30 minute intervals and prior responded surveys. This dataset is available publicly and was obtained by the Irish Commission for Energy Regulation (CER) in an electricity costumer behaviour trial. The data is stored and distributed by the Irish Social Science Data Archive (ISSDA) [30].

The mean hourly household load for the different seasons is shown in Fig. 2. The profiles follow the typical residential dynamic with a small peak in the morning and a larger one at the end of the afternoon. As expected, the mean consumption in winter presents the highest values. The income and education distribution of survey respondents is shown in Fig. 3 and Fig. 4. These distributions show that the used data encompasses different demographics and the results of this experiment are significant for a big part of the Irish population. The survey responses used in the presented analysis result in the following variables: age, employment, social class, internet, adults, children, house type, construction age, bedrooms, electric heating, solar heating, electric water heating, tumble dryer, dishwasher, standalone freezer, water pump, TVs, desktop PCs, laptop PCs, education and income. The value used after processing are continuous integers for all the variables (e.g. higher values represent higher income and education groups) except for the heating, tumble dryer, dishwasher, freezer and pump where they are binary variables indicating the presence of the appliance in the household. The dataset was pre-processed in order to only consider households for which all the variables in analysis have valid values, resulting in a total of 1972 households.

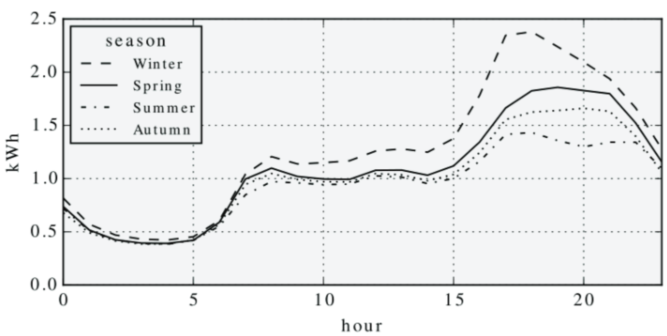


Figure 2 – Hourly aggregated mean seasonal load.

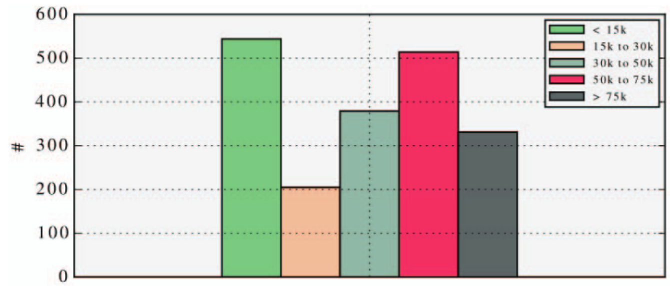


Figure 3 – Income distribution in Euro (€).

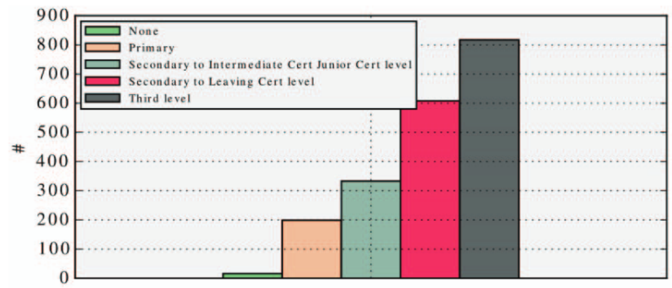


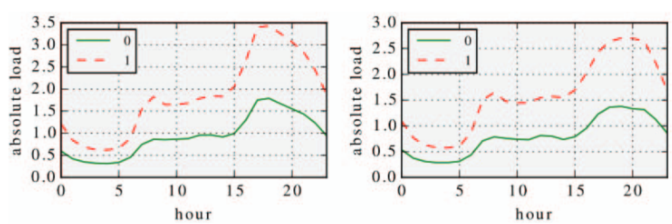
Figure 4 – Survey respondent education distribution.

B. Segmentation

The consumption data of the costumers was filtered by season and aggregated by absolute mean, resulting in a 24-hour mean absolute consumption for each consumer for each season.

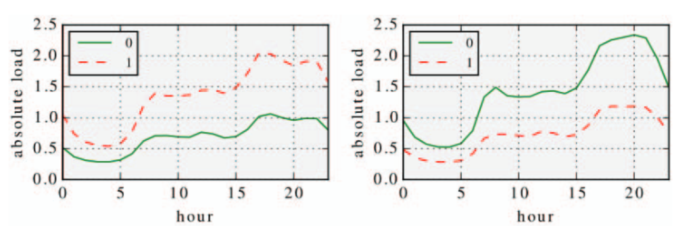
Following the presented methodology, for each season, the number of clusters was tested for values between 2 and 7 and the mean of each one of the CVIs for 10 repeated clusterings was obtained. The number of clusters was selected by majority vote based on the CVI means. For all the seasons the best number of clusters selected is equal to 2.

The obtained cluster centres for each season are shown in Fig. 5 and in Fig. 6. The cluster centres present a high consumption difference between them and so are referred to from now on as "high" and "low" consumption segments.



(a) Winter (b) Spring

Figure 5 – Seasonal clusters: mean load profiles for winter and spring.



(a) Summer (b) Autumn

Figure 6 – Seasonal clusters: mean load profiles for summer and autumn.

C. Feature Analysis

Using the survey based variables as explanatory variables and the cluster identifier 0 or 1 as the dependent variable, probit regression models were fitted to the data in order to find significant correlations between the household characteristics and the consumption profile. The regression results for the winter and summer profiles is shown in Table I and in Table II. The variable lines for which a significant correlation was found are bolded (significance level $\alpha = 0.05$, p-value $< \alpha$).

The winter profiles unveil that the age of the survey respondent, employment status, social class and education are correlated with the "low" consumption profile (negatively correlated to the "high" consumption profile). The usage of internet, number of adults, children, number of bedrooms, usage of tumble dryer, dishwasher, standalone freezer, electric water heating and water pump were found to be positively correlated with the "high" consumption profile.

For the summer profiles a correlation with the "low" consumption profile was found for the age, employment, social class, house type, number of laptops and education. The usage of internet, number of adults, children, usage of tumble dryer, dishwasher, standalone freezer, water pump and electric water heating is correlated to the "high" consumption profile.

The high correlation of number of adults, children and usage of high energy intensive appliances with higher consumptions follows common sense. The fact that social class, respondent employment status and education are correlated with "low" consumption is interesting taking into account that no significant correlation was found for the income. The correlation of number of laptops with "low" consumption is peculiar but may be based on the low consumption of these devices in comparison to desktop computers.

TABLE I. PROBIT REGRESSION RESULTS FOR WINTER PROFILES

Explanatory variable	Coefficient	Std. error	z	p>|z|	95% Conf. Interval	
Age	**-0.13**	**0.03**	**-4.23**	**0.00**	**-0.20**	**-0.07**
Employment	**-0.51**	**0.09**	**-5.53**	**0.00**	**-0.69**	**-0.33**
Social class	**-0.30**	**0.03**	**-9.69**	**0.00**	**-0.37**	**-0.24**
Internet	**0.31**	**0.10**	**3.29**	**0.00**	**0.13**	**0.50**
Adults	**0.27**	**0.04**	**7.24**	**0.00**	**0.20**	**0.35**
Children	**0.32**	**0.04**	**9.03**	**0.00**	**0.25**	**0.39**
House type	-0.01	0.03	-0.31	0.76	-0.06	0.04
Construction age	0.00	0.00	1.79	0.07	0.00	0.00
Bedrooms	**0.11**	**0.05**	**2.30**	**0.02**	**0.02**	**0.20**
Electric heating	0.14	0.13	1.10	0.27	-0.11	0.38
Solar heating	0.13	0.46	0.28	0.78	-0.77	1.03
Electric w. heating	**0.17**	**0.06**	**2.65**	**0.01**	**0.04**	**0.29**
Solar w. heating	0.00	0.29	0.00	1.00	-0.56	0.56
Tumble dryer	**0.33**	**0.08**	**4.01**	**0.00**	**0.17**	**0.49**
Dishwasher	**0.35**	**0.08**	**4.30**	**0.00**	**0.19**	**0.51**
Standalone freezer	**0.14**	**0.07**	**2.20**	**0.03**	**0.02**	**0.27**
Water pump	**0.18**	**0.08**	**2.25**	**0.02**	**0.02**	**0.33**
TVs	0.01	0.03	0.51	0.61	-0.04	0.07
Desktop PCs	0.06	0.06	0.94	0.35	-0.06	0.17
Laptop PCs	-0.05	0.05	-1.05	0.30	-0.14	0.04
Education	**-0.28**	**0.03**	**-8.60**	**0.00**	**-0.35**	**-0.22**
Income	0.03	0.03	1.13	0.26	-0.02	0.09

TABLE II. PROBIT REGRESSION RESULTS FOR SUMMER PROFILES

Explanatory variable	Coefficient	Std. error	z	p>|z|	95% Conf. Interval	
Age	**-0.17**	**0.03**	**-5.32**	**0.00**	**-0.23**	**-0.11**
Employment	**-0.39**	**0.09**	**-4.32**	**0.00**	**-0.57**	**-0.21**
Social class	**-0.23**	**0.03**	**-7.56**	**0.00**	**-0.29**	**-0.17**
Internet	**0.29**	**0.09**	**3.13**	**0.00**	**0.11**	**0.48**
Adults	**0.38**	**0.04**	**9.78**	**0.00**	**0.30**	**0.45**
Children	**0.27**	**0.04**	**7.70**	**0.00**	**0.20**	**0.34**
House type	**-0.06**	**0.03**	**-2.13**	**0.03**	**-0.11**	**0.00**
Construction age	0.00	0.00	1.69	0.09	0.00	0.00
Bedrooms	0.04	0.05	0.82	0.41	-0.05	0.13
Electric heating	-0.02	0.13	-0.17	0.87	-0.27	0.23
Solar heating	0.11	0.48	0.23	0.82	-0.82	1.05
Electric w. heating	**0.26**	**0.06**	**4.17**	**0.00**	**0.14**	**0.39**
Solar w. heating	-0.39	0.29	-1.33	0.19	-0.96	0.19
Tumble dryer	**0.23**	**0.08**	**2.83**	**0.01**	**0.07**	**0.38**
Dishwasher	**0.35**	**0.08**	**4.42**	**0.00**	**0.20**	**0.51**
Standalone freezer	**0.31**	**0.07**	**4.72**	**0.00**	**0.18**	**0.43**
Water pump	**0.18**	**0.08**	**2.27**	**0.02**	**0.02**	**0.33**
TVs	0.04	0.03	1.57	0.12	-0.01	0.10
Dektop PCs	0.10	0.06	1.75	0.08	-0.01	0.22
Laptop PCs	**-0.10**	**0.05**	**-2.28**	**0.02**	**-0.19**	**-0.01**
Education	**-0.28**	**0.03**	**-8.55**	**0.00**	**-0.34**	**-0.21**
Income	0.02	0.03	0.58	0.56	-0.04	0.07

The regression results indicate that a strong correlation between certain household variables and the consumption profile exist, which means that the prediction of the profile segments based on these variables may be feasible.

D. Prediction Models

The prediction of the consumption profile segments based on the survey data was tested using the three types of models presented in the methodology and the results are obtained using three-fold cross validation. The models are iteratively trained with 2/3 of the data and tested with the remaining 1/3.

For this application, the initial results revealed that the use of a linear SVM model resulted in better results than using a kernel to work in a transformed feature space. The parameters of the models were obtained for each fold using grid search with the following parameter ranges: Linear SVM – C = {1, 2, 4}, tolerance = $\{1 \times 10^{-3}, 1 \times 10^{-2}\}$; Gradient Tree Boosting – Nr. Estimators = {10, 20, 40, 100, 200}, max. depth = {1, 2, 3, 5}, learning rate = {0.5, 1.0, 2.0}.

In a binary classification task the true positive (TP) and false positive (FP) are the number of consumers correctly and incorrectly identified to segment "1" and the true negative (TN) and false negative (FN) are the consumers correctly and incorrectly identified to segment "0". The accuracy and balanced accuracy are given by (1) and (2). The area under the curve (AUC) is equal to the area under the receiver operating (ROC) curve of the classifiers [31].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$\begin{aligned} Balanced\ Accuracy &= 0.5 \times TPR + 0.5 \times TNR \\ &= 0.5\left(\frac{TP}{TP + FN}\right) + 0.5\left(\frac{TN}{FP + TN}\right) \end{aligned} \tag{2}$$

The prediction results are shown in Table III. The SVM algorithm obtained the best scores across the table closely followed by the GBT. A maximum accuracy of 75.8% was obtained for the SVM model for winter in comparison to 74.8% of the GBT and 73.5% of the probit regression.

TABLE III. Seasonal And Mean Results Of The Prediction Models. The Results Are Obtained Using Three-Fold Cross Validation

Season	Model	Accuracy %	Balanced Accuracy %	AUC %
Winter	probit	73.5	71.1	76.4
	SVM	75.8	73.6	80.2
	GBT	74.8	72.4	78.7
Spring	probit	73.2	69.9	75
	SVM	74.8	72.1	78.4
	GBT	74.1	71.7	77.2
Summer	probit	72.2	69.9	75.7
	SVM	75.1	73.5	80.0
	GBT	73.9	72.6	78.9
Autumn	probit	71.4	71.0	76.6
	SVM	74.5	73.5	80.4
	GBT	73.7	73.3	79.6
Mean	**probit**	**72.6**	**70.5**	**75.9**
	SVM	**75.1**	**73.2**	**79.8**
	GBT	**74.2**	**72.5**	**78.6**

V. Conclusions

This paper presents a new approach to predict consumer consumption profile segments based on household characteristics. Statistical and machine learning models were developed to predict the demand profile of the consumers. SVM models presented the highest accuracy for this application, achieving of up to 75.8%.

References

[1] P. R. S. Jota, V. R. B. Silva, and F. G. Jota, "Building load management using cluster and statistical analyses," *Int. J. Electr. Power Energy Syst.*, vol. 33, no. 8, pp. 1498–1505, Oct. 2011.

[2] I. Benítez, A. Quijano, J.-L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *Int. J. Electr. Power Energy Syst.*, vol. 55, pp. 437–448, Feb. 2014.

[3] M. Misiti, Y. Misiti, and G. Oppenheim, "Optimized clusters for disaggregated electricity load forecasting," *REVSTAT*, vol. 8, pp. 105–124, 2010.

[4] F. M. Andersen, H. V. Larsen, and T. K. Boomsma, "Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers," *Energy Convers. Manag.*, vol. 68, pp. 244–252, Apr. 2013.

[5] G. Chicco and I. Ilie, "Support vector clustering of electrical load pattern data," *Power Syst. IEEE Trans.*, vol. 24, no. 3, pp. 1619–1628, 2009.

[6] N. Mahmoudi-Kohan, M. P. Moghaddam, and M. K. Sheikh-El-Eslami, "An annual framework for clustering-based pricing for an electricity retailer," *Electr. Power Syst. Res.*, vol. 80, no. 9, pp. 1042–1048, Sep. 2010.

[7] J. J. López, J. a. Aguado, F. Martín, F. Muñoz, a. Rodríguez, and J. E. Ruiz, "Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers," *Electr. Power Syst. Res.*, vol. 81, no. 2, pp. 716–724, Feb. 2011.

[8] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 596–602, May 2005.

[9] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, and M. Kolehmainen, "Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data," *Appl. Energy*, vol. 87, no. 11, pp. 3538–3545, Nov. 2010.

[10] L. Hernández, C. Baladrón, J. Aguiar, B. Carro, and A. Sánchez-Esguevillas, "Classification and Clustering of Electricity Demand Patterns in Industrial Parks," *Energies*, vol. 5, no. 12, pp. 5215–5228, Dec. 2012.

[11] F. Rodrigues, J. Duarte, and V. Figueiredo, "A comparative analysis of clustering algorithms applied to load profiling," in *Proceedings of the 3rd International Conference on Machine Learning and Data Mining*, 2003, pp. 73–85.

[12] S. Valero, M. Ortiz, and C. Senabre, "Application of Self-Organizing Maps for Classification and Filtering of Electrical Customer Load Patterns," *Proc. 2004 Conf. Power Energy Syst.*, 2004.

[13] X. Zhang and C. Sun, "Dynamic intelligent cleaning model of dirty electric load data," *Energy Convers. Manag.*, vol. 49, no. 4, pp. 564–569, Apr. 2008.

[14] O. Dzobo, K. Alvehag, C. T. Gaunt, and R. Herman, "Multi-dimensional customer segmentation model for power system reliability-worth analysis," *Int. J. Electr. Power Energy Syst.*, vol. 62, pp. 532–539, Nov. 2014.

[15] A. Mutanen and M. Ruska, "Customer classification and load profiling method for distribution systems," *Power Deliv. IEEE …*, vol. 26, no. 3, pp. 1755–1763, 2011.

[16] T. F. Sanquist, H. Orr, B. Shui, and A. C. Bittner, "Lifestyle factors in U.S. residential electricity consumption," *Energy Policy*, vol. 42, pp. 354–364, Mar. 2012.

[17] A. Kavousian, R. Rajagopal, and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy*, vol. 55, pp. 184–194, Jun. 2013.

[18] M. Bedir, E. Hasselaar, and L. Itard, "Determinants of electricity consumption in Dutch dwellings," *Energy Build.*, vol. 58, pp. 194–207, Mar. 2013.

[19] B. Sütterlin, T. a. Brunner, and M. Siegrist, "Who puts the most energy into energy conservation? A segmentation of energy consumers based on energy-related behavioral characteristics," *Energy Policy*, vol. 39, no. 12, pp. 8137–8152, Dec. 2011.

[20] F. Fusco, M. Wurst, and J. Yoon, "Mining residential household information from low-resolution smart meter data," in *Proceedings of the 21st International Conference on Pattern Recognition*, 2012, pp. 3545–3548.

[21] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, pp. 397–410, Dec. 2014.

[22] T. Wijaya, T. Ganu, and D. Chakraborty, "Consumer segmentation and knowledge extraction from smart meter and survey data," in *Proceedings of the 2014 International COnference on Data Mining*, 2014, pp. 226–234.

[23] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, "Clustering analysis of residential electricity demand profiles," *Appl. Energy*, vol. 135, pp. 461–471, Dec. 2014.

[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. fifth Berkeley Symp. …*, vol. 1, no. 14, pp. 281–297, 1967.

[25] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, Jan. 1973.

[26] D. L. Davies, D. W. Bouldin, and Other, "A cluster separation measure.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, 1979.

[27] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[28] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2013.

[29] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 297, pp. 273–297, 1995.

[30] ISSDA, "Data from the Commission for Energy Regulation." [Online]. Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[31] J. a Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, no. 4, pp. 29–36, 1982.