16th World Congress of the International Fuzzy Systems Association (IFSA)
9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)

Fuzzy clustering and prediction of electricity demand based on household characteristics

Joaquim L. Viegas[1] **Susana M. Vieira**[1] **João M. C. Sousa**[1]

[1]IDMEC, LAETA, Instituto Superior Técnico, Universidade de Lisboa, Portugal
[1]{joaquim.viegas, susana.vieira, jmsousa}@tecnico.ulisboa.pt

Abstract

The electricity market has been significantly changing in the last decade. The deployment of smart meters is enabling the logging of huge amounts of data relating to the operations of utilities with the potential of being translated into valuable knowledge on the behaviour of consumers. This work proposes a methodology for predicting the typical daily load profile of electricity usage based on static data using fuzzy clustering and modelling. The methodology intends to: (1) determine consumer segments based on the metering data using the fuzzy c-means clustering algorithm, and (2) develop Takagi-Sugeno fuzzy models in order to predict the demand profile of the consumers.

Keywords: Fuzzy clustering, Fuzzy inference system, Smart meter data, Household energy consumption.

1. Introduction

Significant 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. Utilities are becoming very invested in the research and application of new technologies, wishing to achieve higher efficiency and lower losses [1].

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 consumption of costumers. 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 electricity customers and desired high sampling frequencies in smart metering, huge volumes 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 [2, 3]. This paper proposes a methodology for: (1) the segmentation of residential electricity consumers, and (2) the prediction of electricity demand profiles based on household characteristics. A fuzzy modelling approach is used with the intent of obtaining transparent and interpretable models with higher accuracy than classical regression models.

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 low.

2. Related work

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 demand side management strategies (DSM) [4], [5], load forecasting [6], [7], tariff setting [8, 9, 10], marketing and bad data detection. The clustering methods found to be used are mostly the K-means algorithm [5, 11, 12, 13, 14]. Fuzzy clustering [15] has 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.

The use of static data related to household characteristics (e.g. income, inhabitants, education and construction year) 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, 17, 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 novelty of this work in comparison to

the existing literature is the focus on the prediction of electricity consumption profiles based on household characteristics, not only on the analysis of the relationship between them, and the application of fuzzy clustering and modelling in this type of application.

3. Methodology

The methodology is pictured in figure 1. The metering data is processed, aggregated and filtered according to context in order to obtain the typical load profiles (TLP). Following, the load profiles are automatically segmented and Probit regression, fuzzy and support vector machines (SVM) models are derived in order to predict the segments of the consumers based on their household characteristics. The regression and fuzzy models are also used for knowledge extraction in order to correlate the segments to the survey data.

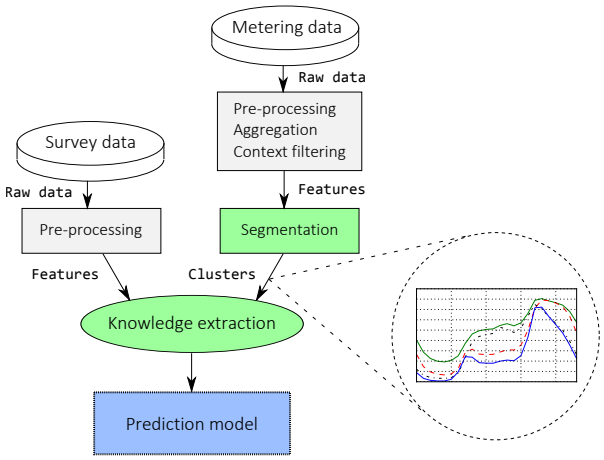


Figure 1: Proposed methodology.

3.1. Typical load profile extraction

The TLP extraction consist in the pre-processing, aggregation and context filtering of the metering data. The pre-processing consists in processing any missing values and is detailed in the dataset presentation in section 4. The context filtering consists in selecting the data of a specific season, consumer group or type of day. The aggregation consists in generating a 24 hour profile from the context filtered data for each consumer (e.g. absolute mean, normalized mean and variance).

3.2. Segmentation

The fuzzy c-means (FCM) [24] algorithm is used in this methodology. In comparison to the classical crisp clustering algorithms like K-Means the samples are not member of an unique cluster and instead are characterized with a membership degree to each one of the clusters. Similarly to the K-Means algorithm FCM also as problems related to the determination of the number of cluster and initial cluster centers.

In order to automatically select the best number of clusters different cluster validity indices (CVIs) were used to test multiple configurations. A majority vote is used to select the best number of clusters.

3.3. Modelling

In order to predict the segment of the consumption profile of a household three different models were used: Probit regression, Takagi-Sugeno fuzzy inference system (TS-FIS) [25] and support vector machines (SVM) for comparison. The regression and fuzzy models are also used in order to extract knowledge on the relationship between inputs and outputs.

3.3.1. Probit regression

Based on the work of Rhodes et al. [23], the relationship between the demand profiles and survey variables is studied 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.

3.3.2. Takagi-Sugeno fuzzy inference system

Fuzzy models are "grey box" and transparent models that allow the approximation of non-linear systems with no previous knowledge of the system to be modelled. Fuzzy inference systems have the advantage, in comparison to other non-linear modelling techniques, to not only provide transparency but also linguistic interpretation in the form of rules.

In this work, TS-FIS are derived from the data. These consist in fuzzy rules where each rule describes a local input-output relation. With TS-FIS, each discriminant function consists, for the binary classification case, in rules of the type

$$R_i: \text{ If } x_1 \text{is } A_{i1} \text{ and } ... \text{ and } x_M \text{ is } A_{iM}$$
$$\text{then } d_i(\mathbf{x}) = f_i(\mathbf{x}), i = 1, 2, ..., K \qquad (1)$$

where f_i is the consequent function of rule R_i. The output of the discriminant function $d_i(\mathbf{x})$ can be interpreted as a score (or evidence) for the the positive example given the input feature vector $\mathbf{x}$. The degree of activation of the ith rule is given by $\beta_i = \prod_{j=1}^{M} \mu_{A_{ij}}(\mathbf{x})$, where $\mu_{A_{ij}}(\mathbf{x}) : \mathbb{R} \rightarrow [0, 1]$.The discriminant output is computed by aggregating the individual rules contributions: $d(\mathbf{x}) = \frac{\sum_{i=1}^{K} \beta_i f_i(\mathbf{x})}{\sum_{i=1}^{K} \beta_i}$.

A sample $\mathbf{x}$ is considered positive if the score is higher than a certain γ threshold $d_i(\mathbf{x}) > \gamma$.

The number of rules K and the antecedent fuzzy sets A_{ij} are determined by fuzzy clustering in the product space of the input variables. FCM is used to determine the cluster centres and the number of clusters was determined through cross-validation. The consequent functions $f_i(\mathbf{x})$ are linear functions determined by ordinary-least squares (OLS) in the space of the input and output variables.

3.3.3. Support Vector Machines

SVM [26] 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.

4. Experimental results

This section presents the dataset used to test the presented methodology and the segmentation, knowledge extraction and prediction results.

4.1. Dataset

The presented methodology was tested using the data from 4232 Irish households over a period of 1.5 years consisting in electricity consumption at 30-minute intervals and prior responded surveys. This dataset is available publicly and was obtained by the Irish CER (Commission for Energy Regulation) in an electricity costumer behaviour trial. The data is stored and distributed by the ISSDA (Irish Social Science Data Archive) [27].

Fig. 2 presents the mean hourly household load for the different seasons. 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. Fig. 3 and 4 present the income and education distribution of survey respondents. These distributions show that the used data encompasses different demographics.

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 water pump where they are binary variables indicating the presence of the appliance in the household.

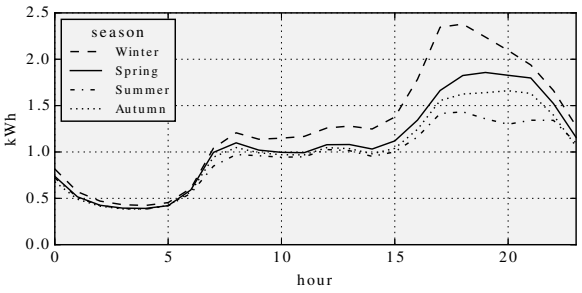


Figure 2: Hourly aggregated mean seasonal load

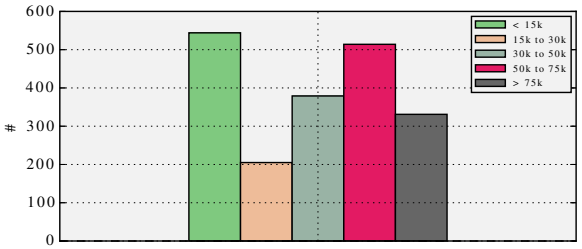


Figure 3: Income distribution in Euro

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 which have no missing values for the survey variables used and metering data.

4.2. 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 and the full year profile. Following the presented methodology, for each type of profile, 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 and the full year the best number of clusters selected is equal to 2. Figures 5, 6 and 7 present the obtained cluster centres for the full year and each season. The cluster centres present a high consump-

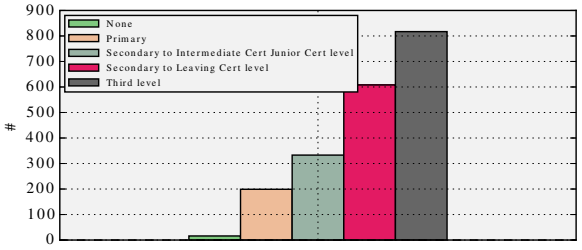


Figure 4: Survey respondent education distribution

tion difference between them and so are referred to from now on as "high" and "low" consumption segments.

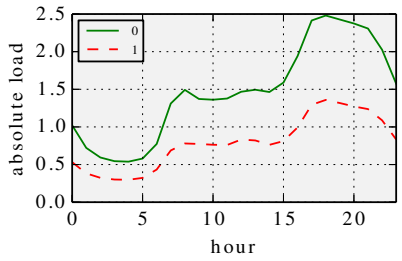
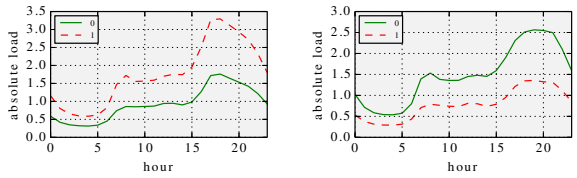
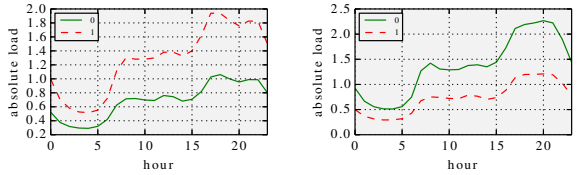


Figure 5: Full year absolute clusters: mean load profiles.



(a) Winter (b) Spring

Figure 6: Seasonal absolute clusters: mean load profiles for winter and spring



(a) Summer (b) Autumn

Figure 7: Seasonal absolute clusters: mean load profiles for summer and autumn

4.3. Knowledge extraction

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. Table 1 present the regression results for the full year profiles. The variable lines for which a significant correlation was found are bolded (significance level $\alpha = 0.05$, p-value $< \alpha$).

It was found that the age of the survey respondent, employment status, social class and education are correlated with the "low" consumption profile. The usage of internet, number of adults, children, number of bedrooms, electric water heating, usage of tumble dryer, dishwasher and standalone freezer were found to be correlated with the "high" consumption profile (negatively correlated to the "low" 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 results of the regression for the seasonal profiles presented very similar results.

Regarding the fuzzy modelling approach for the full year profiles Figure 8 presents the derived membership functions using the parameters which resulted in maximum accuracy. For the majority of the variables the functions are overlapping. Only for the construction age a clear separation between them is visible, separating newer and older aged habitations. The consequent function coefficients are presented in Table 1, in which the variables that were found to more strongly correlate with an "high" consumption profile are the number of adults, children, solar heating, electric heating, tumble dryer, dishwasher and standalone freezer. Solar water heating was the variable with higher correlation with the "low" consumption profile.

4.4. Prediction models

The prediction of the consumption profile segments based on the survey data was done using the three types of models presented in the methodology and the results are obtained using 10-fold cross validation.

For this application initial results revealed the use of a linear SVM generated better results than using a kernel such as the radial basis function type. 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 $= \{1 \times 10^{-3}, 1 \times 10^{-2}\}$; TS-FIS - $K = \{2, 3, 4\}$.

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 obtained following equations 2 and 3. The area under the curve (AUC) is equal to the area under the receiver operating (ROC) curve of the classifiers [28].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$\begin{aligned} Balanced\ accuracy &= 0.5 \times TPR + 0.5 \times TNR \\ &= 0.5\frac{TP}{TP + FN} + 0.5\frac{TN}{FP + TN} \end{aligned} \tag{3}$$

The prediction results are presented in table 2. The SVM algorithm obtained the best results in general but the TS-FIS achieved very close results

Probit regression							TS-FIS consequent functions	
Explanatory variable	Coeff.	Std. error	z	p>|z|	95% Conf.	Interval	$f_1(\mathbf{x})$ coeff.	$f_2(\mathbf{x})$ coeff.
Age	**0.14**	**0.03**	**4.37**	**0.00**	**0.08**	**0.20**	-0.019	-0.041
Employment	**0.37**	**0.09**	**4.17**	**0.00**	**0.20**	**0.55**	-0.02	-0.03
Social class	**0.26**	**0.03**	**8.63**	**0.00**	**0.20**	**0.33**	-0.017	0.052
Internet	**-0.27**	**0.09**	**-2.92**	**0.00**	**-0.45**	**-0.09**	-0.037	-0.099
Adults	**-0.32**	**0.04**	**-8.31**	**0.00**	**-0.39**	**-0.24**	-0.095	-0.116
Children	**-0.32**	**0.04**	**-8.96**	**0.00**	**-0.39**	**-0.25**	-0.143	-0.152
House type	0.05	0.03	1.82	0.07	0.00	0.10	-0.014	-0.073
Construction age	0.00	0.00	-1.50	0.14	0.00	0.00	-0.002	-0.000
Bedrooms	-0.06	0.05	-1.42	0.16	-0.15	0.02	-0.070	-0.033
Electric heating	-0.08	0.12	-0.67	0.50	-0.32	0.16	-0.016	-0.118
Solar heating	-0.28	0.46	-0.62	0.54	-1181.00	0.62	-0.160	-0.042
Electric w. heating	**-0.28**	**0.06**	**-4.40**	**0.00**	**-0.40**	**-0.15**	-0.075	-0.061
Solar w. heating	0.34	0.29	1.18	0.24	-0.22	0.90	0.104	-0.081
Tumble dryer	**-0.32**	**0.08**	**-4.11**	**0.00**	**-0.48**	**-0.17**	-0.106	-0.169
Dishwasher	**-0.42**	**0.08**	**-5.37**	**0.00**	**-0.58**	**-0.27**	-0.157	0.021
Standalone freezer	**-0.27**	**0.07**	**-4.25**	**0.00**	**-0.40**	**-0.15**	-0.014	-0.102
Water pump	-0.13	0.08	-1.73	0.08	-0.29	0.02	-0.020	-0.043
TVs	-0.03	0.03	-1.17	0.24	-0.08	0.02	-0.042	-0.004
Dektop PCs	-0.06	0.06	-1.02	0.31	-0.17	0.06	-0.062	-0.023
Laptop PCs	0.08	0.05	1.77	0.08	-0.01	0.17	-0.023	0.038
Education	**0.26**	**0.03**	**7.97**	**0.00**	**0.19**	**0.32**	0.012	-0.009
Income	0.00	0.03	-0.06	0.95	-0.06	0.05	-0.003	-0.007

Table 1: Probit regression results and TS-FIS consequent function parameters for the full year profiles.

Season	Model	% Acc.	% Acc. balanced	% AUC
Winter	probit	73.2	71.8	76.6
	SVM	75.9	74.9	80.5
	TS-FIS	75.0	74.1	80.1
Spring	probit	73.5	71.5	76.0
	SVM	74.9	74.0	79.2
	TS-FIS	75.4	73.8	79.5
Summer	probit	73.3	72.4	76.7
	SVM	75.4	74.7	80.5
	TS-FIS	73.4	72.3	77.3
Autumn	probit	72.8	72.3	76.4
	SVM	75.0	74.0	79.6
	TS-FIS	72.3	71.6	76.4
Full year	probit	72.8	72.0	76.9
	SVM	76.1	75.9	80.4
	TS-FIS	72.8	72.0	76.8

Table 2: Seasonal and full year results of the prediction models. The results are obtained using 10-fold cross-validation.

for the winter and spring seasons. The TS-FIS has problems with the seasons of summer autumn and full year prediction.

All the developed models obtained accuracies of over 70%. Table 1 presents the coefficients of the consequent linear functions of the TS-FIS using two clusters and Figure 8 presents the membership functions.

The non-separated membership functions for all the variables except the construction age may indicate the use of variables that are affecting negatively the performance of the model indicating the need for feature selection, which is intended in future work.

5. Conclusions

This paper presents a transparent fuzzy clustering and modelling based methodology which is able to predict consumer electricity consumption profiles based on survey data. The SVM models achieved the best performance but lack the transparency of the TS-FIS and ability to interpret coefficients. Based on the developed models a utility company could be able to classify its new costumers into "low" or "high" consumption segments and based on this information advise them on tariff choices and energy saving solutions.

The comparison of the seasonal and full year prediction results show that more research should go into the context filtering of the profiles in order to obtain better fuzzy modelling performance related to the seasons of winter and summer.

The membership functions presented in figure 8 and TS-FIS coefficients of table 1 may indicate the need of feature selection in order to reduce the number of variables (as only the construction age variable is resulting in separated membership functions).

Acknowledgement

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

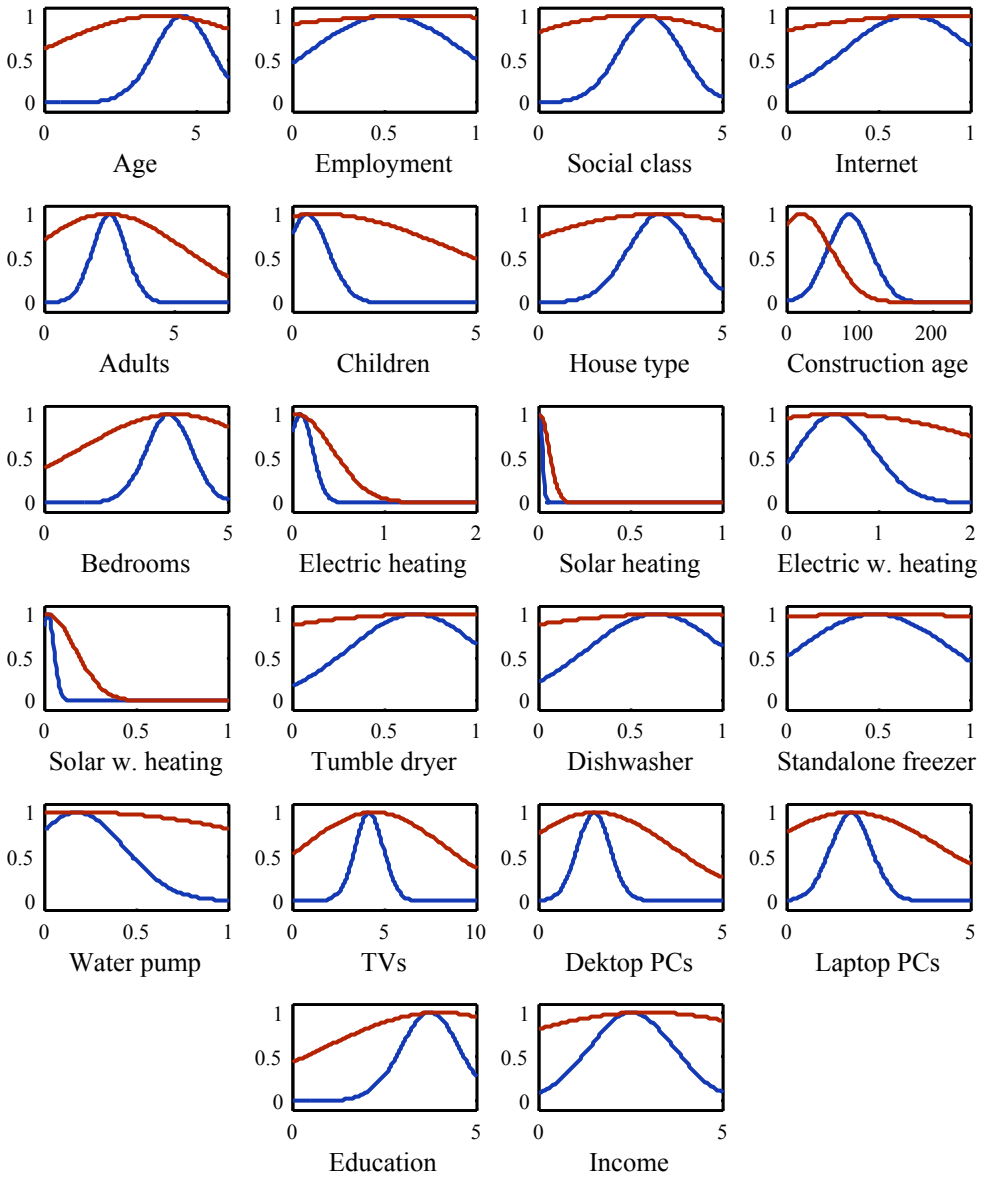


Figure 8: Input membership functions obtained by FCM clustering.

Potential (POPH).

References

[1] Marian Hayn, Valentin Bertsch, and Wolf Fichtner. Electricity load profiles in Europe: The importance of household segmentation. *Energy Research & Social Science*, 3:30–45, September 2014.

[2] Kai-le Zhou, Shan-lin Yang, and Chao Shen. A review of electric load classification in smart grid environment. *Renewable and Sustainable Energy Reviews*, 24:103–110, August 2013.

[3] Cynthia Rudin and David Waltz. Machine learning for the New York City power grid. *Pattern Analysis and ...*, 34(2):328–345, 2012.

[4] Patricia R.S. Jota, Valéria R.B. Silva, and Fábio G. Jota. Building load management using cluster and statistical analyses. *International Journal of Electrical Power & Energy Systems*, 33(8):1498–1505, October 2011.

[5] Ignacio Benítez, Alfredo Quijano, José-Luis Díez, and Ignacio Delgado. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. *International Journal of Electrical Power & Energy Systems*, 55:437–448, February 2014.

[6] Michel Misiti, Yves Misiti, and Georges Oppenheim. Optimized clusters for disaggregated electricity load forecasting. *REVSTAT - Statistical Journal*, 8(2):105–124, 2010.

[7] 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 Conversion and Management*, 68:244–252, April 2013.

[8] G. Chicco and I. S. Ilie. Support vector clustering of electrical load pattern data. *Power Sys-*

tems, IEEE Transactions on, 24(3):1619–1628, 2009.

[9] N. Mahmoudi-Kohan, M. Parsa Moghaddam, and M.K. Sheikh-El-Eslami. An annual framework for clustering-based pricing for an electricity retailer. *Electric Power Systems Research*, 80(9):1042–1048, September 2010.

[10] José J. López, José a. Aguado, F. Martín, F. Muñoz, a. Rodríguez, and José E. Ruiz. Hopfield-K-Means clustering algorithm: A proposal for the segmentation of electricity customers. *Electric Power Systems Research*, 81(2):716–724, February 2011.

[11] V. Figueiredo, F. Rodrigues, Z. Vale, and J.B. Gouveia. An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques. *IEEE Transactions on Power Systems*, 20(2):596–602, May 2005.

[12] Teemu Räsänen, Dimitrios Voukantsis, Harri Niska, Kostas Karatzas, and Mikko Kolehmainen. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy*, 87(11):3538–3545, November 2010.

[13] Luis Hernández, Carlos Baladrón, Javier Aguiar, Belén Carro, and Antonio Sánchez-Esguevillas. Classification and Clustering of Electricity Demand Patterns in Industrial Parks. *Energies*, 5(12):5215–5228, December 2012.

[14] Fátima Rodrigues, Jorge Duarte, and Vera Figueiredo. A comparative analysis of clustering algorithms applied to load profiling. *Machine Learning and ...*, pages 73–85, 2003.

[15] Xiaoxing Zhang and Caixin Sun. Dynamic intelligent cleaning model of dirty electric load data. *Energy Conversion and Management*, 49(4):564–569, April 2008.

[16] Thomas F. Sanquist, Heather Orr, Bin Shui, and Alvah C. Bittner. Lifestyle factors in U.S. residential electricity consumption. *Energy Policy*, 42:354–364, March 2012.

[17] Amir Kavousian, Ram Rajagopal, and Martin Fischer. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, 55:184–194, June 2013.

[18] Merve Bedir, Evert Hasselaar, and Laure Itard. Determinants of electricity consumption in Dutch dwellings. *Energy and Buildings*, 58:194–207, March 2013.

[19] Bernadette Sütterlin, Thomas a. Brunner, and Michael Siegrist. Who puts the most energy into energy conservation? A segmentation of energy consumers based on energy-related behavioral characteristics. *Energy Policy*, 39(12):8137–8152, December 2011.

[20] Francesco Fusco, Michael Wurst, and JW Yoon. Mining residential household information from low-resolution smart meter data. *Pattern Recognition (ICPR), 2012 ...*, (Icpr):3545–3548, 2012.

[21] Christian Beckel, Leyna Sadamori, Thorsten Staake, and Silvia Santini. Revealing household characteristics from smart meter data. *Energy*, 78:397–410, December 2014.

[22] T. K. Wijaya, T. Ganu, and D. Chakraborty. Consumer segmentation and knowledge extraction from smart meter and survey data. *... Conference on Data ...*, pages 226–234, 2014.

[23] Joshua D. Rhodes, Wesley J. Cole, Charles R. Upshaw, Thomas F. Edgar, and Michael E. Webber. Clustering analysis of residential electricity demand profiles. *Applied Energy*, 135:461–471, December 2014.

[24] J. C. Bezdek, R. Ehrlich, and W. Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203, 1984.

[25] Tomohiro Takagi and Michio Sugeno. Fuzzy Identification of Systems and Its Application to Modeling and Control, 1985.

[26] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 297:273–297, 1995.

[27] ISSDA. Data from the Commission for Energy Regulation - www.ucd.ie/issda.

[28] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(4):29–36, 1982.