Working Paper: Predicting Energy Customer Vulnerability Using Smart Meter Data. [☆]

Anastasia Ushakova

University College London

Slava Mikhaylov

University of Essex

Abstract

Supporting vulnerable consumers and reducing fuel poverty are major priorities for policy makers in the energy sector. With the availability of streaming data from smart meters we are able to develop simple and reliable methods of identifying vulnerable energy customers and as a result develop targeted policy interventions. This study investigates how vulnerable customers can be identified from natural gas consumption data. Neural networks, random forest, naive Bayes, and support vector machines were assessed for classification of consumer vulnerability. Random forest, with the prediction accuracy of 94.6 percent, outperforms other prediction models. Our study provides additional evidence that machine learning methods can be deployed by policymakers and insights teams to predict vulnerability from patterns of consumer behaviour.

Keywords: `energy customer vulnerability, prediction models, consumer behaviour, smart meters`

2019 MSC:

[☆] This work is funded by the UK ESRC Consumer Data Research Centre (CDRC) grant reference ES/L011840/1. The data that was used in the study is available via application from CDRC.

[*] Corresponding: Anastasia Ushakova

Email addresses: `anastasia.ushakova@ed.ac.uk` (Anastasia Ushakova), `s.mikhaylov@essex.ac.uk` (Slava Mikhaylov)

Highlights

- The paper surveys various classification methods to predict energy customers vulnerability in the UK residential sector using smart meter data
- Random forest have shown to have the best prediction performance for smart meter data classification
- Morning and evening peak hours have shown to be significant for the distinction between vulnerable and non-vulnerable class
- Smart meter data at half hourly resolution shows the potential for prediction of various customer groups

1. Introduction

The UK was the first European country to introduce in 1994 a policy of energy suppliers' obligation in energy efficiency. Initially, the focus of the policy was on meeting the targets of carbon savings at household level. This was expected to be achieved through subsidised installation of low carbon efficiency measures in UK homes. This obligation, while moderate in the beginning, has become a major component of climate change policy and was further expanded to tackle fuel poverty and ensure affordable access to energy for all UK citizens. Such targets are imposed through Energy Company Obligation (ECO) and administered by the government regulator, Office of Gas and Electricity Markets (OFGEM). Apart from the UK, policies to tackle fuel poverty poverty have been introduced in New Zealand [1, 2], Indonesia [3], Japan [4] and also, in a number of European Union countries such as Italy, France, Belgium and Spain.

For major domestic energy suppliers, compliance with these regulations is of huge importance and companies support several programs to provide house insulation as well as financial support for those who may struggle to pay energy bills. Fuel poverty (or vulnerability) in this sense is identified on the basis of limited ability to access energy at home. In other words, customers may be considered vulnerable if they consume less than they would have had support been provided. For this paper, we narrow our focus to those people who may have difficulty in terms of the financial cost that adequate energy provision requires. A major challenge for energy suppliers in the UK is to identify potentially vulnerable consumers who have not yet self-selected into any of the available support schemes. This paper proposes an approach that could be used to segment and identify potentially vulnerable customers so as to enrol them into available support programs. We propose a prediction model that identifies vulnerable customers based on patterns of energy consumption. We utilise the data on gas consumption from smart meters installed across northern England and Scotland for the period from 2014 to January 2015, and existing data on household energy vulnerability.

Previous research showed the feasibility of using machine learning methods to identify target populations for energy companies' marketing campaigns [5, 6, 7]. Furthermore, smart meter data has previously been used to forecast energy demand [8, 9]. However, this paper is the first use of smart meter data for targeted classification of customers. This study aims to show the possibility of various machine learning methods to help answer a challenging public policy question — how we can identify vulnerable energy customers using data on their consumption available from smart meters.

The paper is structured as follows. The second section reviews the relevant literature on smart meter research and machine learning methods that have been employed to analyse smart meter data. This is followed by an overview of energy suppliers' obligation in the UK, fuel poverty and vulnerable customer issues. The third and fourth sections address data and methodology and outline the analysis strategy that was used to classify customers and discover different consumption patterns in the data. The last section concludes and provides suggestions for further research and the policy implications of our findings.

2. Previous work

Research fields that investigate energy consumption range from engineering and informatics to economics and political science. Such a wide range of disciplines is associated with the complexity of investigation of energy consumption behaviour. Furthermore, there is a requirement for the development of research methods that can be applied to new and varied data types. Literature on smart meters and related issues of energy efficiency for demand-side management has evolved considerably over the last ten years. Expected impacts on efficiency are related to reducing consumer spending on electricity and gas that contributes to overall reduction in carbon emissions. Stromback et al. [10] provides an overview of how smart meters could contribute to energy efficiency goals in Europe. They further stressed the importance of clear customer segmentation as well as active customer participation through provision of feedback. Issues of fuel poverty and energy customers' vulnerability were extensively discussed in [11, 12, 13, 14, 15, 16]. Rosenow et al .[17] provides a critical assessment of fuel poverty defined by policymakers in the UK, including the issues of measuring fuel poverty based on household income and property characteristics. They conclude that performance of existing indicators is not entirely satisfactory and argue for the development of better measures of fuel poverty.

2.1. Modelling energy consumption behaviour

De Silva et al. [18] points out that conventional statistical methods such as linear regression, for example, may face a number of limitations once applied to streaming data available on electricity consumption. As a response, various machine learning techniques applied to smart meter data are considered and implemented in research in this field. For example, Chicco [5] provides a useful summary and comparison of clustering techniques applied to smart meter data for the grouping of customers, which could be further used to target different types of customers for tariff modification and subsequently contribute to meeting energy efficiency goals. One of the most common suggestions is segmentation of customers by type of activity and commercial characteristics. This approach was further considered in Beckel et al. [7] where the authors study data from smart meters for 4,232 households in Ireland over 1.5 years. This work demonstrates the feasibility of performing combined supervised machine learning and multiple regression analysis to reveal customer characteristics, their model achieved a 70% accuracy level. The work of [19] and Sanchez et al. [20], which looked at half-hourly electricity smart meter data, provided a solid foundation for the analysis of overall trends among different customers by using Fourier analysis, self-organizing mappings and various clustering methods. Results have shown clear customer segmentation based on different types such as high usage customers, low usage customers, business customers, and minimal users. This highlighted the feasibility of observing expected trends in peak hours and showed a clear differentiation in consumption patterns among segmented types. Dong et al. [21], used peak time classification to define customers from

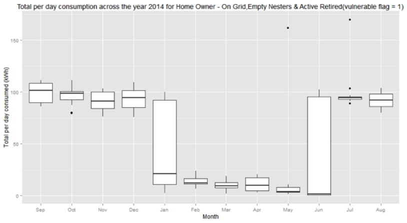

Figure 1: Vulnerable *“Retired and Empty Nester”* customer

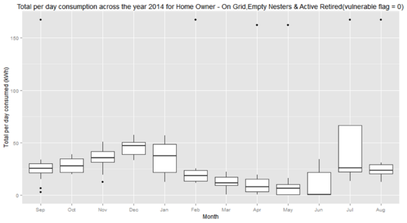

Figure 2: Non-vulnerable *“Retired and Empty Nester”* customer

4,000 Irish households who are most suitable for energy campaigns. . Kwac et al. [6] extended such analysis by using different feature extraction that served as a base for segmentation of customers by lifestyle and consumption behaviour.

Silipo and Winters [22] used electricity smart meter data from 6,000 Irish households and business recordings to provide a reliable prediction model of power shortages and surpluses as well as contribute to targeting mechanisms for finding customers who could be subject to different contract offers. Similar to McDonald et al. [19] the authors showed the effects of weekly and 24 hour seasonality. Most of the clusters showed significant differences in consumption during weekdays and weekends as well as for mornings and evenings. This approach was further expanded in Oates [23] and Liao [24] who demonstrate the application of various time-series clustering techniques to smart meter data. Haben et al. [25] have taken this approach even further by attempting to find a measure of robustness of the clustering methods when applied to smart meter data.

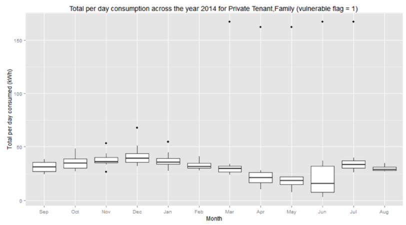

Figure 3: Non-vulnerable *"Family"* customer

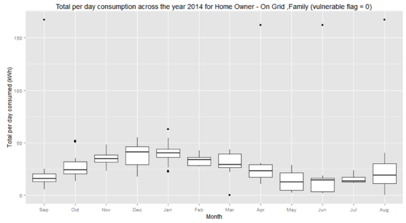

Figure 4: Vulnerable *"Family"* customer

3. Patterns of energy consumption

Data for the analysis was sourced from a large energy supplier in the UK[1] and divided into the following categories – gas consumption and vulnerability characteristics. Data on vulnerability is placed in a separate group as it would be more convenient to have a separate matched sample on vulnerable customers and then compare it with those individuals who have vulnerability flags as absent but may potentially be vulnerable. This is explored further in the methodology section.

The study was based on the sample of 1,919 smart meters from a region in northern England and parts of Scotland. Each meter was recording half-hourly consumption in kWh or Wh depending on the source. For the sake of simplicity, a binary vulnerability flag was created to indicate whether one or more support measures associated vulnerability characteristics were applied to the customer.[2]

[1] We cannot disclose the name of the company due to NDA.

[2] This includes information on customers who have been enrolled in priority services, belong

Data	N (smart meters)	N (days)	N (daily readings)	N (total observations)
One month sample	1,919	28	48	2,372,592
Overall dataset	1,919	390	48	33,309,120

Table 1: *Data structure.* The structure of our smart meter database and the one month sample we are using in the prediction model below.

For the analysis below, we use all smart meter readings for the month of February that we then use later in our prediction model. Table 1 provides an overview of our data. In our sample 24% of customers were classified as vulnerable and 76% as non-vulnerable. On average, per day, overall consumption is recorded at 98.69 kWh (median 67.44) with standard deviation 39.34 (minimum at 0 and maximum at 181.72).

As a simple visualisation of our data we present below monthly consumption patterns for two groups of customers as categorised by the energy supplier. Figures 1-4 illustrate an example of consumption patterns for retired consumers and families, both in the case of vulnerable and non-vulnerable consumers. Overall, retired customers receiving support tend to consume more gas compared to non-vulnerable customers. The family group in Figures 3 and **??** demonstrate consistent consumption over the year, and in fact increased energy consumption during spring time in comparison to previous figures. Presence of the vulnerability flag does not necessarily imply a sizeable difference in consumption for sampled individuals.

Figures 1-4 are based on data of very fine granularity and used here to highlight the complexity of aggregating the data on consumption due to differences in individual consumption profiles. Many additional customer characteristics are omitted, and we cannot assume that sampled customers are representative of consumption patterns for a given life-stage group. Tenancy and property

to a group of credit customers in debt, customers on Fuel direct, customers receiving a grant, vulnerable customers off supply and those who receive a warm home discount. We plan to explore these multiple characteristics in future work.

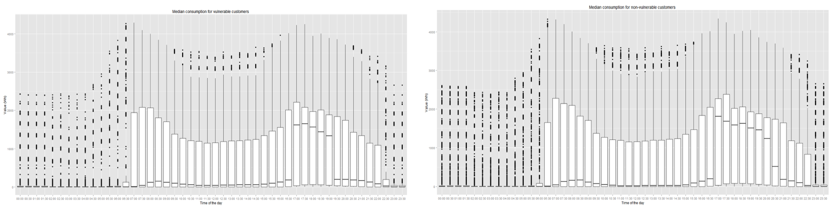

Figure 5: Median half-hourly consumption for vulnerable and non-vulnerable consumers for the month of February.

characteristics as well as geographical location may play a significant role in the observed differences in consumption.

As a further illustration we look at all individuals for weekdays in one month. Figure 5 plots median consumption by vulnerable and non-vulnerable customers during February. It is clear, that the median half-hourly consumption for vulnerable consumers exhibits similar peaks to non-vulnerable consumers, however, there is a slight difference in outliers and magnitude of peaks. We selected February for both visualisation and prediction with the underlying assumption that there is greater variation in the consumption patterns throughout the winter. For Scotland and Northern England, according to Met Office [26] January and February tend to be the coldest months of the year, but February may be more isolated from the effect of winter holidays [21].

The results in Figure 5 indicate high levels of variability within the data, as well as presence of outliers especially for consumption during night hours. While most of the vulnerable sample tend not to use gas during night, there are still a number of individuals exhibiting high consumption levels. It is also very difficult to discern any differences in the consumption patterns between two groups of customers.

4. Predicting consumer vulnerability

The methodology for this study is based on supervised machine learning techniques that are commonly applied in big data analytics and were previously used in the analysis of smart meter data. Nevertheless, the analysis in this paper extends existing literature with more targeted classification and prediction — identification of vulnerable customers.

Least squares and k nearest neighbours models are generally considered the simplest approaches for prediction and classification yet are associated with some statistical instability, particularly in sparse data [27]. As a consequence, we compare a number of alternative methods such as random forest, neural networks, support vector machines, and naive Bayes. As random forest outperformed during our analysis, we will dedicate most of the description to this methodology.

Our motivation for initially utilising a random forest method is primarily driven by growing acknowledgment of its performance for prediction and classification analysis [28]. The model simplifies analysis as it does not require narrowing of the sample or exclusion of certain days since it automatically incorporates patterns in the data corresponding to distinctive individuals. Previously, researchers tended to omit weekends or holiday periods from the analysis as they are often associated with greater heterogeneity among customers [29]. Instead, as our prediction model learns from these 'outlier' patterns, we keep all days so as to allow the models to have better precision in identifying vulnerable customers.

4.1. Random forest

Random forest classification is performed in the following stages. First, the algorithm selects a bootstrap sample to be analysed. The tree is then built through repetitive steps until the optimal combination of variables for predictions with minimal error is found. Each time, the model selects variables at random. In our case, we have 48 variables that correspond to each half-hour smart meter reading every day of the year. The outcome variable is the vulnerability flag for each consumer. The learning algorithm begins on two randomly-selected predictors and expands until covering all 48 predictors. The tree is identical to the decision tree mechanism, where the decision is based on how each variable contributes to further splitting of the data until we can reach our final classification split – into vulnerable and non-vulnerable classes. One advantage of using a random forest model is that it allows for the building multiple trees, rather than just one. Through such a process we mainly look for trees that would build associations between input and output variables. A higher variation in the data allows the algorithm to easily differentiate what contributes to vulnerable and non-vulnerable classes, and split the data further. We do not need to correct the model for seasonality or time dependencies as random forest logically would separate those in the training stage. A brief overview of the methodology is given below:

- The input variables sequence is represented by the sequence $b = \{1...B\}$
- A bootstrap sample is drawn from the training set and random forest tress are built until the minimum size of input variables necessary for training is reached.
- The output of this process is represented by a set of trees $\{T_b\}_{b=1}^{B}$
- The class is then predicted for new or unseen data through the majority vote

  $$\hat{C}_{rf}^{B}(x) = \text{majority vote}\{\hat{C}_b(x)\}_{b=1}^{B} \ , \tag{1}$$

  where $\hat{C}_b(x)$ is the classification decision of the bth tree.

As we observed from the data visualisation, for smart meter data we would expect evening or morning gas consumption levels to have a greater impact on the learning process, while overnight or afternoon consumption should have a relatively smaller influence on the relationship between input and outcome variables.

One of the advantages of using random forest models is low probability of over fitting the data [27] as it is mainly based on decision trees rather than optimisation problems. The optimisation nature of the algorithms are core to the neural network and support vector machine algorithms. As part of the robustness studies, we include these models and briefly discuss them below.

4.2. Neural networks, support vector machines and naive Bayes

The models discussed here are suggested for the analysis of large datasets [27]. The specific choice of the model is often motivated by data variation and whether we expect a linear or nonlinear relationship between predictors and the outcome.

Neural network methodology is based on defining neurons that connect input variables to the outcome, the multilayer structure of the model allows it to represent complex non-linear mappings. In our specification, the analysis is built on a logistic regression model for the hidden layer that connects smart meter readings to a binary vulnerability flag. Minimisation of the sum of squared errors is done by the gradient descent algorithm.

Gradient descent works by using the first order condition of the function in order to find the local minimum point. By taking small steps from a proxy of gradient for a given function, both local maximum or minimum points can be approached through a number of iterations. For this study the number of iterations was raised depending on the size of the sample due to the fact that each customer has a unique combination of inputs and the model may need a reasonable amount of time to converge. Neural networks have been previously been used for energy consumption point prediction [30, 31, 9, 8]. Neural network models usually outperform other approaches such as linear regression, decision trees or support vector machines for the point prediction using historical data. However, in our case we observe a rather poor performance, perhaps due to the classification nature of our prediction problem and noisiness of the data. The latter issue complicates finding a unique solution to the optimisation problem.

Support vector machines (SVMs) are based on the minimization of the cost function through a similar gradient descent approach. Instead of having a hidden layer connecting the input and outcome variable, the algorithm is based on initially creating a nonlinear feature space where it then seeks to fit a linear regression that may separate the features into two classes. While the use of SVMs in energy consumption studies is not extensive, several studies show good prediction performance of such models. For example, Mohandes et al. [32] focus on wind speed prediction from historic daily averages using multi-layer perceptron (MLP) neural networks and support vector machines. Support vector machines outperformed MLP in terms of prediction accuracy. Dong et al. [33] use SVMs to predict energy consumption of commercial buildings in Singapore.

Finally, we also considered a naive Bayes classifier, as in our setting it allows is to calculate the probability of a users vulnerability flag by forming a posterior about the outcome. This posterior updates as more smart meter readings are taken. Thus, with more data available we would expect greater prediction accuracy. The probability of the outcome variable to be either zero or one is estimated using the maximum likelihood approach. Naive Bayes, as shown in Rish [34], relies mainly on the assumption that the features are independent of the predicted class, and performs well on highly-interdependent features. The prediction power would gradually decrease if the class zero is over represented in the sample. In our case, after re-balancing the sample, we could not report a highly visible difference in the prediction power using naive Bayes. This is likely attributable to high variation in our half-hour loads. In addition, the heterogeneous levels of interdependencies associated with half-hour consumption may also arise from idiosyncratic usage of natural gas at household level.

In practical work, algorithms often differ in how they utilise predictors that are less statistically important for identifying the relationship between input and outcome [35]. Whereas random forest models benefit from such weak inputs, for neural network and support vector machines this additional noise may detrimentally affect the solution. Caruana and Niculescu-mizil [36] show that for highly variable and complex data sets, or those that have information on real-world complex problems, naive Bayes is expected to be outperformed by models like random forest. Our results confirm this earlier prediction.

5. Results

We select the optimal model parameters through ten-fold cross validation. To assess the performance of the models, Table 2 reports accuracy, precision,

Model	Accuracy	Recall	Precision	F-score
Neural Network	60.11%	0.64	0.66	0.65
SVM (radial kernel)	76.2%	0.81	0.99	0.89
Naive Bayes	56.0%	0.81	0.85	0.83
Random Forest	**94.6%**	**0.81**	**0.79**	**0.80**

Table 2: Results (ten folds cross validation) for each model that was used to predict vulnerability flag using consumption data.

recall, and F-score. Our results suggest that Random Forest outperforms alternative models in terms of overall accuracy. We suggest that random forest may have a greater power in differentiating similar patterns of consumption.

As the choice of alternative models was largely driven by their popularity in academic research, our results are in line with the literature in related fields. For example, Lines et al.[37] focus on appliance consumption predictions and compare naive Bayes, random forest, neural network, and SVM (also using cross-validation to select optimal model parameters). They show that Random Forest slightly outperforms other models on their data.

One of the advantages of using random forest is its interpretability. Alternative models like neural network and SVMs are often treated as “black-box” solutions. With random forest we can assess which variables are significant for prediction accuracy. Figure 6 provides a summary of variable importance tables that indicate the variables in the order of importance for prediction power and their contribution to subsequent tree splits based on Gini impurity criterion. By importance ranking, as expected, morning and evening peak hours have a strong contribution to prediction accuracy. By Gini, morning hours tend to be more important in their contribution to the split of the decision tree.

In line with the original argument by Breiman [35], we believe that weak inputs in the dataset make achieving high prediction accuracy with neural network or SVM more challenging. Variation that arises from these variables adds more noise and contributes to more confusion in convergence to local minimum point.

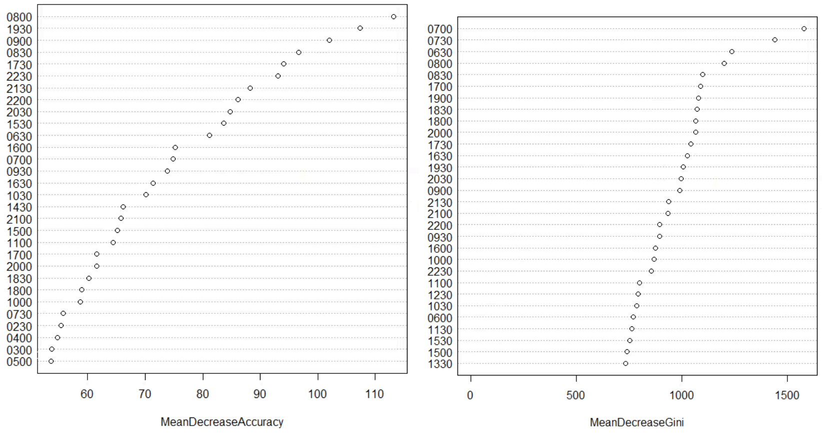

Figure 6: Mean Decrease Accuracy and Gini by variable importance

In our case, as we see from the variables' importance in Figure 6, almost half of the variables are not critical for prediction accuracy. Random forest appears to have taken weakness/importance into account, thus achieving maximum noise reduction. Lastly, based on law of large numbers, with the increasing number of trees there is likely convergence to a unique solution.

6. Conclusions, Limitations and Future Work

The research presented above aimed to answer the question of whether there is the potential to identify vulnerable natural gas customers by using data from smart meters using various machine learning methods. Vulnerable customers were expected to under-consume in transition to the winter period, yet this was hardly observed in our data. Vulnerable customers have shown more constant and distributed over the day consumption profiles while the non-vulnerable tend to exhibit peaks and have quite uniform patterns of consumption within the sample, implying that they may leave home at certain periods while vulnerable customers may consistently use gas in their homes. Nevertheless, some patterns in both groups were similar which may have been a reason as to why most of the models failed to provide high prediction accuracy. In our prediction model, random forest produces better prediction accuracy compared to alternative models.

It is important to acknowledge that while this paper produces some insights, it should be seen as a demonstration, rather than complete solution, for how smart meter data can be used to understand and predict vulnerability vulnerability. Further research may consider using predictive mechanisms on segmented customer groups obtained from various clustering methods; such a suggestion was also made by Haghi and Toole [9]. It may also be fruitful to train models on larger samples and consider weather conditions as an additional input variable — this would enable the achievement of more precision in the training of the model by giving it associations with weather parameters. Prediction models may also be performed on features that are based on household and property characteristics. Such extensive analysis may provide a better precision and accuracy when predicting vulnerability.

Further limitations are related to defining vulnerability and benchmarking. In this paper vulnerability was assessed based on the available data on customers that were enrolled in support services. Conceptually, this measure requires further investigation, perhaps with a comparison to alternative measures. In addition, customers who received support may change their consumption behaviour, thus it may be a strong assumption that non-vulnerable customers exhibiting patterns similar to those present in the vulnerable sample may also be vulnerable. To control for this it may be valuable to observe the indicators on the time when support was provided so we could see so-called treatment effect on consumption caused by efficiency measures.

To round up the paper, perhaps the most trivial policy implication to be drawn from this research is a relationship between high heterogeneity of gas consumption households and the meeting of the objectives set by OFGEM. The ECO requirement on identifying vulnerable customers and those at risk of fuel poverty may be a hard task to meet in the absence of a strategy for identifying such customers backed up by high quality data analysis. Thus, policymakers may consider either providing energy companies with the right tools to perform such strategies or soften the current requirement. Furthermore, energy consumption vulnerability remains an ambiguous concept which may require further research to build more inclusive and transparent indicators. As suggested in Schmidt and Weigt [38], the study of energy demand and consumption requires a highly interdisciplinary approach especially if policymakers are interested in shaping and transforming current energy systems. Thus, both social and political science as well as engineering and data science may be helpful in answering such research questions.

References

[1] P. Howden-Chapman, H. Viggers, R. Chapman, K. O?Sullivan, L. T. Barnard, B. Lloyd, Tackling cold housing and fuel poverty in new zealand: A review of policies, research, and health impacts, Energy Policy 49 (2012) 134 – 142, special Section: Fuel Poverty Comes of Age: Commemorating 21 Years of Research and Policy. doi:http://dx.doi.org/10.1016/j.enpol.2011.09.044.

[2] K. C. O'Sullivan, P. L. Howden-Chapman, G. M. Fougere, Fuel poverty, policy, and equity in new zealand: the promise of prepayment metering, Energy Research & Social Science 7 (2015) 99–107.

[3] R. K. Andadari, P. Mulder, P. Rietveld, Energy poverty reduction by fuel switching. impact evaluation of the {LPG} conversion program in indonesia, Energy Policy 66 (2014) 436 – 449. doi:http://dx.doi.org/10.1016/j.enpol.2013.11.021.

[4] S. Okushima, Measuring energy poverty in japan, 2004?2013, Energy Policy 98 (2016) 557 – 564. doi:http://dx.doi.org/10.1016/j.enpol.2016.09.005.

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

[6] J. Kwac, J. Flora, R. Rajagopal, Household energy consumption segmentation using hourly data, Smart Grid, IEEE Transactions on 5 (1) (2014) 420–430.

[7] C. Beckel, L. Sadamori, T. Staake, S. Santini, Revealing household characteristics from smart meter data, Energy 78 (2014) 397–410.

[8] J. Lee, Y.-c. Kim, G.-L. Park, An analysis of smart meter readings using artificial neural networks, Convergence and Hybrid Information Technology (2012) 182–188.

[9] A. Haghi, O. Toole, The use of smart meter data to forecast electricity demand,, CS229 Course paper.

[10] J. Stromback, C. Dromacque, M. H. Yassin, The potential of smart meter enabled programs to increase energy and systems efficiency: a mass pilot comparison short name: Empower demand, VaasaETT, Global Energy Think Tank.

[11] S. Bouzarovski, S. Petrova, S. Tirado-Herrero, From fuel poverty to energy vulnerability: The importance of services, needs and practices, Tech. rep., SPRU-Science and Technology Policy Research, University of Sussex (2014).

[12] J. Hills, Getting the measure of fuel poverty: Final report of the fuel poverty review, Department of Energy and Climate Change (DECC) CASE report 72.

[13] B. Legendre, O. Ricci, Measuring fuel poverty in france: Which households are the most fuel vulnerable?, Energy Economics 49 (2015) 620–628.

[14] L. Middlemiss, R. Gillard, How can you live like that?: energy vulnerability and the dynamic experience of fuel poverty in the uk.

[15] T. Sefton, Targeting fuel poverty in england: is the government getting warm?, Fiscal Studies 23 (3) (2002) 369–399.

[16] B. Boardman, Fixing Fuel Poverty: Challenges and Solutions, Earthscan, 2010.

[17] J. Rosenow, R. Platt, B. Flanagan, Fuel poverty and energy efficiency obligations–a critical assessment of the supplier obligation in the uk, Energy Policy 62 (2013) 1194–1203.

[18] D. De Silva, X. Yu, D. Alahakoon, G. Holmes, A data mining framework for electricity consumption analysis from meter data, Industrial Informatics, IEEE Transactions on 7 (3) (2011) 399–407.

[19] B. McDonald, P. Pudney, J. Rong, Pattern recognition and segmentation of smart meter data, ANZIAM Journal 54 (2014) 105–150.

[20] I. B. Sánchez, I. D. Espinós, L. M. Sarrión, A. Q. López, I. N. Burgos, Clients segmentation according to their domestic energy consumption by the use of self-organizing maps, in: Energy Market, 2009. EMM. 6th International Conference on the European, IEEE, 2009, pp. 1–6.

[21] H.-A. Cao, C. Beckel, T. Staake, Are domestic load profiles stable over time? an attempt to identify target households for demand side management campaigns, in: Industrial Electronics Society, IECON 2013 - 39th Annual Conference of the IEEE, IEEE, 2013, pp. 4733–4738.

[22] R. Silipo, P. Winters, Big data , smart energy , and predictive analytics time series prediction of smart energy data,, Tech. rep., KNIME (2013).

[23] T. Oates, Identifying distinctive subsequences in multivariate time series by clustering, in: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1999, pp. 322–326.

[24] T. W. Liao, Clustering of time series data-a survey, Pattern recognition 38 (11) (2005) 1857–1874.

[25] S. Haben, C. Singleton, P. Grindrod, Analysis and clustering of residential customers energy behavioral demand using smart meter data, IEEE transactions on smart grid 7 (1) (2016) 136–144.

[26] Met Office, Met Office: Climate Summaries, `http://www.metoffice.gov.uk/climate/uk/summaries`, [Accessed: 2015-09-12] (2015).

[27] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, Vol. 1, Springer series in statistics Springer, Berlin, 2001.

[28] M. Weiss, A. Helfenstein, F. Mattern, T. Staake, Leveraging smart meter data to recognize home appliances, in: Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on, IEEE, 2012, pp. 190–197.

[29] D.-W.-I. C. Flath, D.-W.-I. D. Nicolay, T. Conte, P. D. C. van Dinther, L. Filipova-Neumann, Cluster analysis of smart metering data, Business & Information Systems Engineering 4 (1) (2012) 31–39.

[30] S. J. Nizami, A. Z. Al-Garni, Forecasting electric energy consumption using neural networks, Energy Policy 23 (12) (1995) 1097 – 1104. `doi:http://dx.doi.org/10.1016/0301-4215(95)00116-6`.
URL `http://www.sciencedirect.com/science/article/pii/0301421595001166`

[31] G. K. Tso, K. K. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy 32 (9) (2007) 1761 – 1768. `doi:http://dx.doi.org/10.1016/j.energy.2006.11.010`.
URL `http://www.sciencedirect.com/science/article/pii/S0360544206003288`

[32] M. Mohandes, T. Halawani, S. Rehman, A. A. Hussain, Support vector machines for wind speed prediction, Renewable Energy 29 (6) (2004) 939–947.

[33] B. Dong, C. Cao, S. E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings 37 (5) (2005) 545–553.

[34] I. Rish, An empirical study of the naive bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, IBM New York, 2001, pp. 41–46.

[35] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.
URL http://dx.doi.org/10.1023/A:1010933404324

[36] R. Caruana, A. Niculescu-mizil, An empirical comparison of supervised learning algorithms, in: In Proc. 23 rd Intl. Conf. Machine learning (ICML 2006, 2006, pp. 161–168.

[37] J. Lines, A. Bagnall, P. Caiger-Smith, S. Anderson, Classification of Household Devices by Electricity Usage Profiles, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 403–412. doi:10.1007/978-3-642-23878-9_48.

[38] S. Schmidt, H. Weigt, Interdisciplinary energy research and energy consumption: What, why, and how?, Energy Research & Social Science 10 (2015) 206–219.

[33] B. Dong, C. Cao, S. E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings 37 (5) (2005) 545–553.

[34] I. Rish, An empirical study of the naive bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, IBM New York, 2001, pp. 41–46.

[35] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.
URL http://dx.doi.org/10.1023/A:1010933404324

[36] R. Caruana, A. Niculescu-mizil, An empirical comparison of supervised learning algorithms, in: In Proc. 23 rd Intl. Conf. Machine learning (ICML 2006, 2006, pp. 161–168.

[37] J. Lines, A. Bagnall, P. Caiger-Smith, S. Anderson, Classification of Household Devices by Electricity Usage Profiles, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 403–412. doi:10.1007/978-3-642-23878-9_48.

[38] S. Schmidt, H. Weigt, Interdisciplinary energy research and energy consumption: What, why, and how?, Energy Research & Social Science 10 (2015) 206–219.