Classification of new electricity customers based on surveys and smart metering data

Joaquim L. Viegas[a,*], Susana M. Vieira[a], R. Melício[a,b], V. M. F. Mendes[b,c], João M. C. Sousa[a]

[a] *IDMEC, LAETA, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisbon, Portugal*
[b] *Dep.de Física, Escola de Ciências e Tecnologia, Universidade de Évora, Portugal*
[c] *Instituto Superior de Engenharia de Lisboa*

Abstract

This paper proposes a process for the classification of new residential electricity customers. The current state of the art is extended by using a combination of smart metering and survey data and by using model-based feature selection for the classification task. Firstly, the normalized representative consumption profiles of the population are derived through the clustering of data from households. Secondly, new customers are classified using survey data and a limited amount of smart metering data. Thirdly, regression analysis and model-based feature selection results explain the importance of the variables and which are the drivers of different consumption profiles, enabling the extraction of appropriate models. The results of a case study show that the use of survey data significantly increases accuracy of the classification task (up to 20%). Considering four consumption groups, more than half of the customers are correctly classified with only one week of metering data, with more weeks the accuracy is significantly improved. The use of model-based feature selection resulted in the use of a significantly lower number of features allowing an easy interpretation of the derived models.

Keywords:
Data-driven energy efficiency, Electricity customer clustering, Classification of new residential customers, Customer feature selection, Smart metering data, Customer surveys data

*Corresponding author.
Email address: joaquim.viegas@tecnico.ulisboa.pt (Joaquim L. Viegas)

1. Introduction

A game-changing shift has been happening in the utility industry and energy markets. Policy focused on energy efficiency and sustainability is growing fruit of the awareness of current environmental challenges. Liberalization, growing competition between utilities, technological advancements and policy towards a sustainable use of energy sources are pushing utilities to seek innovation and new market related insights.

Electricity is a main energy carrier used around the world for supporting the primary, secondary and tertiary sectors. The commercial and residential energy demand is expected to continue to shift towards electricity and away from primary fuels. By 2040, forecasts indicate that electricity generation will account for more than 40% of global energy consumption and, from 2010 to 2040, global electricity demand is projected to increase by about 85% [1–3].

Technological advancement in the fields of metering, communications and computation are enabling utilities to monitor and save huge amounts of data related to their operation. The deployment of electricity meters with two-way communication capabilities is enabling the logging of the consumption of users with high resolution. The number of advanced metering infrastructure (AMI) installations, also known as smart meters, has surpassed the number of traditional one-way communication meters in the United States [4]. Close to 45 million smart meters are already installed in three Member States (Finland, Italy and Sweden) of the European Union (EU), representing 23 percent of the envisaged installation in the EU by 2020 [5].

The consumption data of customers has the potential to give insights of great importance for utilities and policy makers. Valuable insights can be derived by the knowledge of typical consumption curves of different consumer groups and understanding what are the main drivers of consumption. This knowledge can assist decision makers in the electricity utility industry in developing demand side management (DSM) programs, consumer engagement strategy, marketing, alternative tariff setting methods and demand forecasting tools [6]. Knowledge on the way different demographic groups consume electricity is valuable to study the effect of energy policy on different population groups.

The high number of consumers and desired high sampling frequencies in smart metering implies that huge amounts of data have to be stored and processing grows in complexity. Computational intelligence techniques in the fields of machine learning are starting to be extensively used in order to extract knowledge from the data coming from the grid. These techniques can provide decision makers with predictive models and the ability to extract valuable knowledge.

In order to characterize the behaviour of electricity customers, the clustering of electricity consumption data 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 [7, 8], load forecasting [9–11], tariff setting [12–14], marketing and bad data detection. The clustering methods found to be used are mostly the K-means algorithm [8, 15–18]. Fuzzy clustering [19] 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. Normalization, parametric modelling [10], temperature based normalization [16, 20] and wavelet transformation [9] have been found to be used in the literature.

The use of static data related to household characteristics, e.g., income, number of inhabitants, education, construction year and appliances in relation to static or dynamic energy consumption data is being studied in order to find the main drivers of residential energy consumption. In [21–23] 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 [24] demographic data and psychological and belief related data is studied in comparison to energy consumption. [25, 26] presents studies on the prediction of household information based on smart meter data. In [27, 28] consumptions profiles obtained via clustering are correlated to household characteristics. In [29] a methodology is presented for the characterization of medium voltage electricity customers through clustering and posterior modelling for which the classification of new customers is stated as a possible application.

Classifying new customers is crucial for marketing purposes, as customers with lengthy relationships are less likely to defect and are less affected by new information and offers. Thus, a greater impact of marketing strategies and engagement is expected with new customers [30, 31].

This paper extends the current state of the art by developing a process for the classification of new electricity customers using not only metering data but also using static data on household characteristics. The use of a limited amount of metering data is done in order to emulate the analysis of new electricity customers for which only a small amount of data is available.

The use of model-based feature selection for the discovery of the consumption drivers shows promise in the field.

Based on the clustering of customers' electricity consumption data, the consumption profile of new customers is predicted using survey data and a limited amount of smart metering data. Classification models in combination with model-based and filter feature selection are compared for the classification task, selection and analysis of variables.

The developed process aims to provide an interpretable classification modelling method for the classification of electricity customers and discovery of the drivers of different electricity consumption profiles. The presented results aim to illustrate the application of the proposed process, using data that resulted from smart metering trials encompassing more than three thousand households in Ireland [32]. Requirements for the classification of customers and insights on the drivers of residential electricity consumption are presented.

This paper is organized as follows: Section 2 discusses the uses of the proposed process in the context of the smart grid. Section 3 presents the method for the generation of the populations representative consumption profiles. Section 4 presents the techniques used for modelling, feature selection and model evaluation. Section 5 presents the experimental results and presents the discussion and Section 6 presents the conclusions.

2. Classification of customers in the smart grid

The smart grid is a concept with the purpose of intelligently integrating the generation, transmission and consumption of electricity through technological means [33–37]. A smart electricity grid enables an efficient management of the whole electricity supply chain through innovative applications. The applications can provide the capacity to: securely integrate more renewable energy sources and distributed generation; deliver power in a more efficient and secure manner through advanced control and monitoring; automatically reconfigure the grid to prevent and restore outages; better integrate consumption through DSM; enable consumer engagement in the market [38–41].

Smart metering roll-outs and pilots are paving the way for the development of the smart grid. Meters with two-way communication capabilities are expected to empower consumers by enabling the creation of consumer services and engaging them to actively participate in the electricity market.

In Europe the total investment of smart grids amounted to €3.15 billion in 2014 and smart metering projects account for most of the total investment [38].

The imperative for consumers to be on board is defended in order not only to reap the benefits of a smart grid, but also to make smart metering projects profitable. The extent of the transformation of the grid rests on the needs and the willingness of consumers to pay for the implementation [38, 41]. The right consumers need to be identified, engaged and motivated in order to reap the benefits of smart metering in terms of electricity cost savings, through, e.g., load shifting [42].

Knowledge on the ways electricity is consumed in a population and what are the drivers of consumption dynamics, e.g., demographics, household characteristics and the use of appliances is essential in order to personalize applications, energy services and policy towards a smarter grid.

In the context of the smart grid, the ability to effectively group customers into similar behaviour market segments and to find the segment of new customers is very valuable, e.g., in the following applications:

- Proposing tariff offers or DSM schemes taking into account the expected consumption behaviour of the customers;
- Planning and studying the potential impact of personalized services and offers;
- Offering the energy saving and sustainability services the customers are most likely to be interested in.

The proposed process for clustering and classification of electricity customers enables more effective customer engagement on the part of utilities and smart grid operators. Customer engagement is essential to maximize the willingness of customers to pay for the implementation of this type o grid, either directly or indirectly by increasing the grids efficiency through DSM programs and energy efficiency solutions.

3. Clustering

Clustering methods attempt to group objects based on a definition of similarity. The objective is to find groups of objects with greater similarity between them than to the objects of other groups.

In the scope of this paper and the analysis of customers' representative consumption profiles, clustering methods are used to find which are the groups of customers which have similar consumption curves in some context, e.g., season, type of day. These groups are represented by the populations representative consumption profiles, resulting from aggregating the profile of all the customers of a group, equivalent to the cluster centroid.

The methodology followed to find the customer groups and respective representative consumption profiles is in Figure 1. The clustering process is similar to the one proposed in [29]. Firstly, smart metering data is processed in order to obtain the customers' representative consumption profiles, secondly, various clustering configurations are tested. Configurations are evaluated using multiple clustering validity indexes (CVI) which are used, together with careful visual evaluation, to chose the final configuration and obtain the customer groups and profiles.

3.1. Customers' normalized representative consumption profiles

Smart metering consumption data is composed of a large set of timestamped intervals with consumption values. In order to obtain consumption profiles which can be easily interpreted, visualized and manipulated, the data goes through a process of context filtering, aggregation and pre-processing.

The process of context filtering consists on selecting data which represents a specific context, defined, for example, by a temporal window (e.g. Winter, Summer), type of day (e.g. working day) and location.

Let $\mathbf{x}_i$ be the feature vector (list of variables) associated to customer i. $\mathbf{x}_i = (\mathbf{x}_i^m, \mathbf{x}_i^s)$ where $\mathbf{x}_i^m$ has dimension r equal to the number of variables which characterize a customers representative load profile (LP) or derived load indices (LI) and $\mathbf{x}_i^s$ has dimension t equal to the number of survey variables used. The dimension of a customers feature vector $\mathbf{x_i}$ is $p = r+t$. The LI and survey variables are presented in 5.1 and 5.3. $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\} \subseteq \Re^p$ is the feature dataset of N customers.

After filtering, the consumption data is aggregated in order to reduce the dimension and obtain a curve representative of the whole temporal window. The aggregation is characterized by the period used, e.g., hourly, daily and operator, e.g., mean, median. For example, doing an hourly mean aggregation of the consumption data of customer i will generate a vector $\mathbf{x}_i^m \in \Re^{24}$ in which each element represents the mean consumption in a certain hour.

The final pre-processing consists on the normalization of the data for easier clustering, modelling and representation of different information. This paper focuses on the case of normalization for each customer in which each representative profile is normalized with the maximum value of the profile as normalization factor. The normalization is done with the intent of translating the consumption dynamic in relation to the maximum. This is done in [27–29]. The clustering of absolute representative consumption profiles results, using the same kind of data, on a separation of groups by amount of consumption.Without normalization the different shapes of curves are seemingly overshadowed by the mean absolute consumption while clustering [43].

Figure 2 pictures an example of the clustering results, showing clusters centroids for hourly aggregated absolute and normalized representative profiles. The curves behave in a similar way for different scales in absolute profiles. For normalized consumption profiles the curves are distinct in terms of linearity and consumption between different times of the day.

3.2. K-means clustering

The K-means algorithm [44] is used due to its simplicity, efficiency and scalability. The algorithm has been proven to be adequate for this type of application in the literature [8, 15–18, 45, 46]. Let $\mathbf{S} = \{S_1, ..., S_J\}$ be the groups (sets) of customers clustered together, J the number of clusters and d_e a chosen distance measure. The centroid of a cluster S_k is its mean vector, $\mu_k = \frac{1}{|S_k|}\sum_{\mathbf{x}\in S_k}\mathbf{x}$. The algorithm is an iterative refinement method which, in this application, minimizes the distance between the customers' consumption profiles $\mathbf{x}$ and the populations μ_k, as given by (1).

$$\arg\min_{\mathbf{S}} \sum_{k=1}^{J} \sum_{\mathbf{x}\in S_k} d_e(\mathbf{x}, \mu_k)^2 \tag{1}$$

The difficulty associated with this algorithm is the need to determine the number of clusters and their initial centres. The choice of the number of cluster centres is detailed in the following Section 3.3. The initial cluster centres are generated randomly and the best clustering result of an high number of runs is used.

3.3. Clustering evaluation

A clustering in X is a set of disjoint clusters that partition X into k groups: $\mathbf{S}$ where $\cup_{S_k\in\mathbf{S}} S_k = X, S_k \cap S_l = \emptyset \, \forall \, k \neq l$. The euclidean distance is used and $d_e(\mathbf{x_i}, \mathbf{x_k}) = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{kj})^2}$.

As pictured in Figure 1, multiple CVI are used to evaluate a number of different clustering configurations. If there is no consensus between the different CVI the expert chooses the best configuration based on the analysis of the CVI and visualization of the clustering results.

Three different CVI are used in this work, they evaluate the goodness of the clustering in terms of maximization of inter cluster distances and minimization of intra cluster distances [47].

The Dunn index (D) [48] is a ratio-type index where the cohesion is estimated by the nearest neighbour distance and the separation by the maximum cluster diameter. The original index is defined as,

$$D(\mathbf{S}) = \frac{\min_{S_k \in \mathbf{S}} \{\min_{S_l \in \mathbf{S} \setminus S_k} \{\delta(S_k, S_l)\}\}}{\max_{S_k \in \mathbf{S}} \{\Delta(S_k)\}} \tag{2}$$

where,

$$\delta(S_k, S_l) = \min_{\mathbf{x_i} \in S_k} \min_{\mathbf{x_j} \in S_l} \{d_e(\mathbf{x_i}, \mathbf{x_j})\} \tag{3}$$

$$\Delta(S_k) = \max_{\mathbf{x_i}, \mathbf{x_j} \in S_k} \{d_e(\mathbf{x_i}, \mathbf{x_j})\}. \tag{4}$$

The Davis-Bouldin index (DB) [49] estimates the cohesion based on the distance from the points in a cluster to the centroid and the separation based on the distance between centroids. The DB index is defined as:

$$DB(\mathbf{S}) = \frac{1}{J} \sum_{S_k \in \mathbf{S}} \max_{S_l \in \mathbf{S} \setminus S_k} \left\{ \frac{F(S_k) + F(S_l)}{d_e(\mu_k, \mu_l)} \right\} \tag{5}$$

where,

$$F(S_k) = \frac{1}{|S_k|} \sum_{\mathbf{x_i} \in S_k} d_e(\mathbf{x_i}, \mu_k). \tag{6}$$

The silhouette index (Sil) [50] is a normalized summation-type index. The cohesion is measured based on the distance between all the points in the same cluster and the separation is based on the nearest neighbor distance. The silhouette index is defined as:

$$Sil(\mathbf{S}) = \frac{1}{N} \sum_{S_k \in \mathbf{S}} \sum_{\mathbf{x_i} \in S_k} \frac{b(\mathbf{x_i}, S_k) - a(\mathbf{x_i}, S_k)}{\max\{a(\mathbf{x_i}, S_k), b(\mathbf{x_i}, S_k)\}} \tag{7}$$

where,

$$a(\mathbf{x_i}, S_k) = \frac{1}{|S_k|} \sum_{\mathbf{x_j} \in S_k} d_e(\mathbf{x_i}, \mathbf{x_j}) \tag{8}$$

$$b(\mathbf{x_i}, S_k) = \min_{S_l \in \mathbf{S} \setminus S_k} \Big\{ \frac{1}{|S_l|} \sum_{\mathbf{x_j} \in S_l} d_e(\mathbf{x_i}, \mathbf{x_j}) \Big\}. \tag{9}$$

4. Modelling

4.1. Classification

This work intends to train models to predict the group of a new customer, characterized by a representative consumption profile. Figure 3 pictures the electricity customer classifier.

Features are extracted from the survey responses and smart metering data of the customer. Based on the features the classifier returns a categorical variable y indicative of the customer group in which the customer best fits.

The classifier is a function φ which maps the features of a customer to a categorical variable y, representing one of the J customer groups. It is defined as:

$$\varphi : \Re^p \mapsto y \tag{10}$$

$$y \in \{c_1, c_2, ..., c_J\} \tag{11}$$

Classifiers are trained using the group labels extracted through the clustering of a full year of smart metering data, considered as the ground truth to be inferred from features extracted from a limited amount of smart metering data and survey data.

The two following sections present the modelling approaches used in this methodology.

4.1.1. Logistic regression

The logistic regression (LR) models the posterior probabilities of the J classes via linear function in x while ensuring the sum to one and remaining in $[0, 1]$. The LR model has the form presented in (12), where D represents the input vector [51, 52]. The parameter set of the model is $\theta = \{\beta_{10}, \beta_1^T, ..., \beta_{(J-1)0}, \beta_{(J-1)}^T\}$.

$$\log \frac{\Pr(y=1|D=\mathbf{x})}{\Pr(y=J|D=\mathbf{x})} = \beta_{10} + \beta_1^T \mathbf{x}$$
$$\log \frac{\Pr(y=2|D=\mathbf{x})}{\Pr(y=J|D=\mathbf{x})} = \beta_{20} + \beta_2^T \mathbf{x} \quad (12)$$
$$\vdots$$
$$\log \frac{\Pr(y=J-1|D=\mathbf{x})}{\Pr(y=J|D=\mathbf{x})} = \beta_{(J-1)0} + \beta_{(J-1)}^T \mathbf{x}$$

Using the LR model, if the clustering analysis results in J customer groups, the classifier linearly separates each one of $J-1$ customer groups to the J customer group.

LR is usually fit by maximum likelihood, in the case of the results presented in this paper the Newton-Raphson optimization method is used. For the case of two classes the parameters of the model can be easily interpreted through the significance and sign. In the case of multiple classes the interpretation of the model parameters is more complex due to a total set of $J-1$ parameters for each variable.

The LR model is chosen due to the simplicity (explained by linear functions) and interpretability, enabling the understanding of the role of the different input variables in explaining the outcome [51]. Models with increased complexity, such as artificial neural networks or support vector machines, may provide higher accuracy but lack the transparency of the LR model [53].

4.1.2. Decision trees

Binary decision tree (DT) learning consists on fitting data to a tree-like structure. This type of method partitions the feature space into a set of rectangles and usually fits a constant in each one. This paper makes use of the popular tree-based regression and classification method called CART (Classification And Regression Tree) [51]. Tree-based methods have the advantage of an easy interpretation and can be transformed into a simple set of rules if the number of branches is low.

In order to grow a classification DT the learning algorithm automatically splits the data into two sets at each level, optimizing some criterion which translates the model accuracy. In this paper the Gini index is used, which is a measure of how often a randomly chosen element from the set is incorrectly labelled if it is randomly labelled according to the distribution of labels in the subset. The learning algorithm minimizes the difference of this measure between tree levels through the growth of the DT. Using DT in the multiple class case is straightforward and each end node of the tree will give a probability for the J labels. Figure 4 pictures an example of a partition obtained by binary splitting and corresponding DT.

A classification DT model is chosen, similarly to the LR model, due to its interpretability, providing a popular binary tree representation [51].

4.2. Feature selection

The objective of feature selection (FS) is to choose a subset of the available features by eliminating features with little or no predictive information and also redundant features that are strongly correlated [54]. FS techniques are usually divided into filter, wrapper and embedded methods. Wrapper and embedded are usually referred to as model-based methods and filter techniques as model-free methods.

Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data. Filter techniques are normally easily scalable to very high-dimension datasets and computationally simple, having the disadvantage of not taking into account the interaction with the classifier [55].

Wrapper methods embed the classification model within the feature subset search. The selected set of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm [55].

4.2.1. Regression based filter feature selection

In regression analysis parameters are determined indicating the relationship between the features and the model output. The p-values of the hypothesis tests based on the parameters' standard errors indicate if the corresponding variables are believed to be significantly different from 0 (rejected null hypothesis), thus indicators of the output variable. The regression feature selection method used removes the variables for which the corresponding parameters result in a p-value higher than a certain significance level (5%).

This parametric filter FS technique has been used in multiple studies, together with LR or probit regression, in order to find which are the features which are indicative of a specific electricity consumption profile and are determinants of electricity consumption [22, 23, 28].

4.2.2. Wrapper feature selection

This paper proposes the use of greedy wrapper FS methods to find relations between the characteristics of customers and the typical consumption profile. FS is also done in order to generate interpretable models by significantly reducing the number of features used to classify new customers.

Sequential forward selection and sequential backward elimination [56] are the FS methods used. The forward FS algorithm sequentially selects features, starting with a empty set, choosing the features that improve the most the prediction accuracy. This is done until there is no more improvement in prediction. The backward FS algorithm starts with the full set of features and sequentially removes the ones which result in an improvement in prediction accuracy.

4.3. Model evaluation

In order to maximize the significance of the performance results of the trained classifiers k-fold cross-validation is used [51, 53]. This model validation technique randomly divides the dataset into k folds. The classifier is then trained (using $k-1$ folds) and evaluated (using 1 fold) k times, as pictured in Figure 5. The modelling approach is then evaluated through the mean and standard deviation of the accuracy.

In order to do an unbiased FS the methods presented in Section 4.2 are used only based on the training sets so that the process is totally independent from the test data. The wrapper FS methods also make use of cross-validation to evaluate the feature subsets.

5. Results and discussion

5.1. Dataset

The proposed methodology is applied to data from 4232 Irish households monitored for one and a half year. The dataset consists of electricity consumption data logged at 30 minute intervals and surveys responded before the start of the trial. This dataset resulted from an electricity customer behaviour trial by the Irish Commission for Energy Regulation (CER). The data is stored and maintained by the Irish Social Science Data Archive (ISSDA) [32].

The mean hourly consumption for the four seasons is pictured in Figure 6. Consumption follows the typical residential dynamic with a small peak in the morning and lunch time, a larger one at the end of the afternoon and low consumption during the night. As expected, the mean consumption in winter presents the highest values due to the heating needs.

The distribution of the survey responses on social class and number of children per household is pictured in Figure 7. AB is upper middle class and middle class, C1 is lower middle class, C2 is skilled working class, DE is working and non-working classes and F represents farmers. The distributions show that the used data encompasses different demographic groups and household types.

The survey questions used as features are presented in Table 1 to Table 4, along with a description and possible responses. Table 1 presents the features with information on the respondent, Table 2 is related to the habitation characteristics, Table 3 to the heating systems and Table 4 to the appliances.

Survey variables with no response are considered as 'refused'. The customers not considered in the study are the ones who did not respond to the question indicating the number of adults in the household. The final dataset used contains 3440 electricity customers.

5.2. Clustering

This section presents the results from the extraction of features from the customers smart metering data, transformation in representative profiles and clustering in order to obtain the final populations representative consumption profiles.

5.2.1. Customers' representative consumption profiles

In order to obtain the customers' consumption profiles the parameters used to extract the representative features are:

- Context: Only the smart metering data from working days is used and profiles are extracted seasonally;
- Aggregation: The data is aggregated hourly resulting in twenty-four features ($r = 24$);
- Operator: The operator used is the mean.
- Normalization: The profiles are normalized with regards to each customers maximum hourly consumption.

The final customers' representative consumption profiles are equal to the customer normalized mean hourly consumption in working days. The profiles are obtained for each one of the four seasons.

5.2.2. *Populations representative consumption profiles*

Following the proposed methodology, the best number of clusters is found to be equal to four for the four seasons. Figure 8 pictures the evolution of the three CVI used when generating between two and six clusters for the Winter season. The Silhouette, Dunn and Davis-Bouldin indexes indicate, respectively, that the best number of cluster is two, four and five. In order to choose a number of clusters the partitions are visually analysed as pictured in Figure 9, Figure 10 and Figure 11. The figures present the populations representative consumption profiles (cluster centres) and the customers' representative consumption profiles pertaining to the cluster.

With two clusters, as pictured in Figure 9, many customers have a consumption profile different from the centre, indicating the need for an higher number of clusters. With four clusters, as pictured in Figure 10, the clusters are sufficiently compact having a significant number of customers in each group. With five clusters, as pictured in Figure 11, *Cluster 2* has a low number of customers with profiles showing a low similarity. Based on the visual analysis the number of chosen clusters is equal to four. The same process is used for the other seasons.

The final populations representative consumption profiles are pictured in Figure 12. The population is divided mainly due to the following consumption profile characteristics:

- ***Peakiness***: Relation between peak evening consumption and the consumption throughout the rest of the day. For example: in Winter, clusters 1 and 2 have a much higher difference between peak evening and the rest of the days consumption (high *peakiness*), in comparison to clusters 3 and 4 (low *peakiness*).

- **Decline time**: Time at which the consumption starts to rapidly decline after peak evening consumption. For example: in Spring, clusters 2 and 4 have a late declining consumption (late decline) in comparison to clusters 1 and 3 (early decline), specially cluster 3 that has a very early decline in consumption.

- **Off-peak consumption**: Presence of significant consumption during the off-peak hours (night and early morning) in comparison to the rest of the day. For example: in Autumn, cluster 4 presents a significant consumption during the night hours (high off-peak consumption) in comparison to the clusters 1, 2 and 3 (low off-peak consumption).

Summer presents the most different populations consumption profiles in comparison to the other seasons, as pictured by the the consumption profile of *Cluster 2*. This cluster presents a high amount of variability between customers results in a low mean normalized consumption throughout the day.

Table 5 presents the distribution of customers between the different clusters for each one of the seasons. Asides from the Winter clustering, the customers are approximately uniformly distributed between the four groups.

5.3. Classification of new customers and feature selection

Features extracted from metering data and from conducted surveys are used for the classification of new customers. In order to evaluate the process for the classification of new customers, the metering data is limited to an amount starting from no data to ten weeks of data. Due to the high amount of metering data and desire for interpretable models two types of features extracted from the smart metering data are tested: load profile (LP) and load indices (LI).

The LP features are the ones used in the clustering: in this paper they are the hourly aggregated mean consumption normalized on an individual basis. The features differ from the ones used for clustering due to being derived from a limited amount of smart metering data.

The LI are shape indices derived from the LP, these are proposed in [57] and used for the characterization of medium-voltage customers in [29]. LI are used in this paper with the intention of obtaining models of easier interpretation, explaining what consumption characteristics are the most relevant when comparing customers. The indices are presented in Table 6. i_1 is the load factor, i_2 is the off-peak factor, i_3 is the night impact coefficient, i_4 is the lunch impact coefficient and i_5 is the modulation coefficient at off-peak hours. P_{max}, P_{min}, P_{av} are, respectively, the maximum, minimum and average consumption of the corresponding periods.

Table 7 summarizes the smart metering features used in classification. In the case at least one day of metering data is available, a total of $p = r + t = 24 + 47 = 71$ features are available using the LP as the smart metering features and $p = 5 + 47 = 52$ features are available using the LI.

Table 8 and Table 9 present the mean and standard deviation of the accuracy of the trained classifiers, through 5-fold cross-validation, in the cases of no smart metering data, 1, 4, 8 and 10 weeks of available smart metering data (W). In parentheses the mean number of features selected is presented. The results are presented for the LR and DT models, for each season, and further divided by the use of no FS, the filter FS algorithm and forward FS. Backward FS results in a performance closely similar to the use of no FS. Accuracy was used, instead of measures that can correctly deal with class imbalances, such as the Area Under the ROC Curve (AUC) [58], precision/recall and MCC, due to the multiclass nature of the classification problem and the approximately balanced nature of the classes, inferred from Table 5.

The evolution of the LR classifier performance with a growing number of weeks of metering data for the Winter season is pictured in Figure 13. The figure shows that, when using LP, the classification accuracy always benefits from the use of survey features. The difference between the performance of the classifier with and without survey features grows with the number of available weeks of smart metering data. When using LI the difference is only significant for the case when there is not metering data for which the classification is random because no features are available.

Based on the analysis of the results of Table 8 and Table 9, the use of LP results in an better classification performance, proving that the LI are not able to correctly translate all the information needed to classify the customers.

In general, filter FS results in the best accuracy, reducing significantly the number of features in comparison with not using any FS. Using forward FS resulted in an even greater reduction of the number of features at the cost of a reduction of accuracy.

The following paragraphs present a detailed analysis of the classification and feature selection results for:

1. Winter with no metering data;
2. Spring with one week of metering data transformed in LI;
3. Summer with four weeks of metering data transformed in LP;
4. Autumn with eight weeks of metering data transformed in LP.

For the classification of the Winter profiles without any smart metering data Table 10 presents the variables selected by the filter FS algorithm (regression analysis) and Figure 14 pictures the rate of selection of the variables selected by the forward FS throughout the cross-validation process. A maximum mean accuracy of 39% is achieved with the features selected by filter FS. With the forward FS the number of features is reduced from

16 to 9 and 4, respectively for LR and DT, achieving a better accuracy for DT (37.4% with forward and 36.3% with filter FS) and slightly worst with LR (37.3%). The variables selected by forward FS with LR modelling are mainly `age` and `employment`. `heat_solidfuel`, `tumble_dryer` and `electric_cooker` are also selected in more than half of the cross-validation folds. The variable selected by forward FS with DT modelling is mainly `age`. `heat_electricity_plugin` and `electric_cooker` are also selected in the more than half of the cross-validation folds. The age, employment, type of heating and the use of electric cooking appliances are the features which can be used as indicators to separate customers with different consumption profiles.

For the classification of the Spring profiles with one week of smart metering data, translated by LI, Table 11 presents the variables selected by the filter FS algorithm and Figure 17 pictures the rate of selection of the variables selected by the forward FS throughout the cross-validation process. A maximum mean accuracy of 56.5% is achieved with the features selected by filter FS. With the forward FS the number of features is reduced from 20 to 9 and 5, respectively for LR and DT, achieving slightly worst accuracies. The variables selected by forward FS with LR modelling are mainly the five LI (i_1, ..., i_5) and `washing_machine`. The variables selected by forward FS with DT modelling are mainly three LI (i_1, i_3, i_4), indicating that the load factor, night impact and lunch impact are the LI features which can be used as indicators to separate customers with different consumption profiles.

For the classification of the Summer profiles with four weeks of smart metering data, translated by LP, Table 12 presents the variables selected by the filter FS algorithm and Figure 15 pictures the rate of selection of the variables selected by the forward FS throughout the cross-validation process. A maximum mean accuracy of 73.3% is achieved with the features selected by filter FS. With the forward FS the number of features is reduced from 30 to 16 and 5, respectively for LR and DT, achieving slightly worst accuracies (71.7% and 64.9%). The variables selected by forward FS with LR modelling are mainly multiple LP features ($l_1, l_2, l_7, l_{11}, l_{16}, l_{18}, l_{22}, l_{23}, l_{24}$) and `washing_machine`. The variables selected by forward FS with DT modelling are mainly LP features ($l_2, l_{12}, l_{15}, l_{23}$). The consumption behaviour translated by LP features distributed throughout the day in combination with the number of washing machines in the customers household can be used as indicators to separate customers with different consumption profile.

For the classification of the Autumn profiles with eight weeks of smart metering data, translated by LP, Table 13 presents the variables selected by the filter FS algorithm and Figure 16 pictures the rate of selection of the variables selected by the forward FS throughout the cross-validation process. A maximum mean accuracy of 81.6% is achieved with the features selected by filter FS. With the forward FS the number of features is reduced from 32 to 16 and 8, respectively for LR and DT, achieving worst accuracies (77.9% and 70.4%). The variables selected by forward FS with LR modelling are mainly multiple LP features ($l_8, l_{10}, l_{12}, l_{13}, l_{14}, l_{15}, l_{17}, l_{20}, l_{22}, l_{23}, l_{24}$) and `washing_machine`. The variables selected by forward FS with DT modelling are mainly LP features ($l_2, l_3, l_5, l_{21}, l_{23}$). The consumption behaviour translated by LP features distributed throughout the day in combination with the number of washing machines in the customers household can be used as indicators to separate customers with different consumption profile.

Notice the LR results having a high standard deviation of the accuracy, such as the results for ten weeks of metering data for Winter and Spring with no FS, using LP metering features. These result due the inappropriate convergence of the optimization method for LR training. Using forward FS this problem is avoided.

Based on the results, the five most important variables or questions an utility should ask customers on sign-up are:

1. What is the customer employment status;
2. How old the customer is;
3. How many dishwashers are used in the clients household;
4. How many electric cookers are used in the clients household;
5. How many washing machines are used in the clients household.

6. Conclusions

The integration of smart metering in the power grid enables a detailed analysis of the consumption behaviour of electricity customers. Knowledge on the typical consumption profiles of customers and the main drivers of consumption are extremely valuable for decision makers in the utility industry and policy. The engagement and education of consumers is seen as a key task in order to successfully reap the potential benefits of the smart grid [41]. The daily routines and the social context of consumers needs to be correctly taken into account to efficiently plan and target the correct groups for potential DSM programs and create incentives for consumers to act with regard towards sustainability.

The proposed process is a contribution for enabling the modelling of interpretable classifiers to predict the consumption profile group of new customers using smart metering data and survey responses. It enables the discovery of the drivers of consumption profiles, e.g., which characteristics of customers are able to translate consumption behaviour differences. This can contribute to the better engagement of consumers and development of measures to increase efficiency in the power grid.

An application, based on the data from more than three thousand residential electricity customers from Ireland, shows the viability of the proposed methods. Without any metering data the LR is able to correctly classify up to 39% of the customers which is significantly better than randomly inserting the customer in one of the four customer groups (with four customer groups). With the growth of available smart metering data the simulations show an increase in accuracy achieving up to 60%, 70% and 80% accuracy, respectively, with 1, 4 and 8 weeks of data.

The forward FS results pictured are easily interpreted and resulted in the discovery of the most important features when grouping electricity customers by their representative consumption profile. For the Irish population studied in the paper, information on the representative consumption profile throughout all the day results in the highest classification accuracy. A low number of shape indices is not suitable to accurately classify new electricity customers. The number of washing machines in the customers households is revealed to be a very important feature in the classification task, seemingly being the most influencing feature to the considerable increase of accuracy from the use of survey features added to the smart metering features.

Acknowledgements

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).

References

[1] Exxon Mobil Corporation, The outlook for energy: a view to 2040.

[2] OECD, ICT applications for the smart grid: opportunities and policy implications, OECD Digital Economy Papers (190).

[3] D. S. Markovic, D. Zivkovic, I. Branovic, R. Popovic, D. Cvetkovic, Smart power grid and cloud computing, Renewable and Sustainable Energy Reviews 24 (2013) 566–577.

[4] U.S. Energy Information Administration, Annual electric power industry report, Tech. rep. (2013).
URL http://www.eia.gov/electricity/data/eia861/

[5] Commission Européenne, Benchmarking smart metering deployment in the EU-27 with a focus on electricity (2014).

[6] R. Granell, C. J. Axon, D. C. Wallom, Clustering disaggregated load profiles using a Dirichlet process mixture model, Energy Conversion and Management 92 (2015) 507–516.

[7] P. R. Jota, V. R. Silva, F. G. Jota, Building load management using cluster and statistical analyses, International Journal of Electrical Power & Energy Systems 33 (8) (2011) 1498–1505.

[8] I. Benítez, A. Quijano, J.-L. Díez, I. Delgado, Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers, International Journal of Electrical Power & Energy Systems 55 (2014) 437–448.

[9] M. Misiti, Y. Misiti, G. Oppenheim, Optimized clusters for disaggregated electricity load forecasting, REVSTAT - Statistical Journal 8 (2) (2010) 105–124.

[10] F. Andersen, H. Larsen, T. Boomsma, Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers, Energy Conversion and Management 68 (2013) 244–252.

[11] H. R. Sadeghi Keyno, F. Ghaderi, a. Azade, J. Razmi, Forecasting electricity consumption by clustering data in order to decline the periodic

variable's affects and simplification the pattern, Energy Conversion and Management 50 (3) (2009) 829–836.

[12] G. Chicco, I. S. Ilie, Support vector clustering of electrical load pattern data, IEEE Transactions on Power Systems 24 (3) (2009) 1619–1628.

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

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

[15] V. Figueiredo, F. Rodrigues, Z. Vale, J. Gouveia, An electric energy consumer characterization framework based on data mining techniques, IEEE Transactions on Power Systems 20 (2) (2005) 596–602.

[16] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, M. 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) (2010) 3538–3545.

[17] L. Hernández, C. Baladrón, J. Aguiar, B. Carro, A. Sánchez-Esguevillas, Classification and clustering of electricity demand patterns in industrial parks, Energies 5 (12) (2012) 5215–5228.

[18] F. Rodrigues, J. Duarte, V. Figueiredo, Z. Vale, M. Cordeiro, A comparative analysis of clustering algorithms applied to load profiling, in: Machine Learning and Data Mining in Pattern Recognition, Springer, 2003, pp. 73–85.

[19] X. Zhang, C. Sun, Dynamic intelligent cleaning model of dirty electric load data, Energy Conversion and Management 49 (4) (2008) 564–569. doi:10.1016/j.enconman.2007.08.007.

[20] A. Mutanen, M. Ruska, Customer classification and load profiling method for distribution systems, IEEE Transactions on Power Delivery 26 (3) (2011) 1755–1763.

[21] T. F. Sanquist, H. Orr, B. Shui, A. C. Bittner, Lifestyle factors in U.S. residential electricity consumption, Energy Policy 42 (2012) 354–364.

[22] A. Kavousian, R. Rajagopal, 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 55 (2013) 184–194.

[23] M. Bedir, E. Hasselaar, L. Itard, Determinants of electricity consumption in Dutch dwellings, Energy and Buildings 58 (2013) 194–207.

[24] B. Sütterlin, T. a. Brunner, M. Siegrist, Who puts the most energy into energy conservation? A segmentation of energy consumers based on energy-related behavioral characteristics, Energy Policy 39 (12) (2011) 8137–8152.

[25] F. Fusco, M. Wurst, J. W. Yoon, Mining residential household information from low-resolution smart meter data, in: 21st International Conference on Pattern Recognition (ICPR), IEEE, 2012, pp. 3545–3548.

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

[27] T. K. Wijaya, T. Ganu, D. Chakraborty, K. Aberer, D. P. Seetharam, Consumer segmentation and knowledge extraction from smart meter and survey data, in: SIAM International Conference on Data Mining (SDM14), 2014.

[28] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, M. E. Webber, Clustering analysis of residential electricity demand profiles, Applied Energy 135 (2014) 461–471. doi:10.1016/j.apenergy.2014.08.111.

[29] S. Ramos, J. M. Duarte, F. J. Duarte, Z. Vale, A data-mining-based methodology to support MV electricity customers characterization, Energy and Buildings 91 (2015) 16–25.

[30] R. N. Bolton, A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction, Marketing Science 17 (1) (1998) 45–65. doi:10.1287/mksc.17.1.45.

[31] P. C. Verhoef, Understanding the effect of customer relationship management efforts on customer retention and customer share development, Journal of Marketing 67 (4) (2003) 30–45. doi:10.1509/jmkg.67.4.30.18685.

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

[33] M. Welsch, M. Howells, M. Bazilian, J. DeCarolis, S. Hermann, H. Rogner, Modelling elements of smart grids: enhancing the OSeMOSYS (open source energy modelling system) code, Energy 46 (1) (2012) 337–350.

[34] U.S. Department of Energy, The smart grid: an introduction, Tech. rep. (2008).

[35] International Energy Agency, Technology roadmap: smart grids, Tech. rep. (2011).

[36] ETP SmartGrids, European technology platform smart grids: vision and strategy for Europe's electricity networks of the future, Tech. rep. (2006).

[37] A. Battaglini, J. Lilliestam, A. Haas, A. Patt, Development of supersmart grids for a more efficient utilisation of electricity from renewable sources, Journal of Cleaner Production 17 (10) (2009) 911–918.

[38] C. Felix, M. Ardelean, J. Vasiljevska, A. Mengolini, G. Fulli, E. Amoiralis, M. S. Jiménez, C. Filiou, Smart grid projects outlook 2014, European Commision, JRC Science and Policy Reports.

[39] A. Faruqui, D. Harris, R. Hledik, Unlocking the €53 billion savings from smart meters in the eu: How increasing the adoption of dynamic tariffs could make or break the eu's smart grid investment, Energy Policy 38 (10) (2010) 6222–6231.

[40] A. J. Conejo, J. M. Morales, L. Baringo, Real-time demand response model, Smart Grid, IEEE Transactions on 1 (3) (2010) 236–242.

[41] V. Giordano, F. Gangale, G. Fulli, M. Sánchez, J. Dg, I. Onyeji, A. Colta, I. Papaioannou, A. Mengolini, C. Alecu, T. Ojala, I. Maschio,

Smart grid projects in Europe : lessons learned and current developments, European Commision: JRC Scientific and Policy Reports.

[42] Institute of Communication & Computer Systems of the National Technical University of Athen ICCS-NTUA for the European Commission, Study on cost benefit analysis of Smart Metering Systems in EU Member States - Final Report.

[43] J. L. Viegas, S. M. Vieira, R. Melício, V. M. F. Mendes, J. a. M. C. Sousa, Electricity demand profile prediction based on household characteristics, in: Proceedings of the 12th International Conference on the European Energy Market, 2015.

[44] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA., 1967, pp. 281–297.

[45] G. Chicco, Overview and performance assessment of the clustering methods for electrical load pattern grouping, Energy 42 (1) (2012) 68–80. doi:10.1016/j.energy.2011.12.031.
URL `http://dx.doi.org/10.1016/j.energy.2011.12.031`

[46] T. Warren Liao, Clustering of time series data - A survey, Pattern Recognition 38 (2005) 1857–1874.

[47] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, I. n. Perona, An extensive comparative study of cluster validity indices, Pattern Recognition 46 (1) (2013) 243–256.

[48] C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics 3 (3) (1973) 32–57.

[49] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence (2) (1979) 224–227.

[50] P. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.

[51] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer.[Online book], 2008.

[52] D. C. Montgomery, E. A. Peck, G. Vining, Introduction to Linear Regression Analysis, 5th Edition, Wiley, 2012.

[53] M. R. Berthold, C. Borgelt, F. Höppner, F. Klawonn, Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data, Springer, 2010.

[54] S. M. Vieira, J. a. M. Sousa, T. a. Runkler, Two cooperative ant colonies for feature selection using fuzzy models, Expert Systems with Applications 37 (4) (2010) 2714–2723. doi:10.1016/j.eswa.2009.08.026.

[55] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.

[56] J. Kittler, Feature set search algorithms, Pattern recognition and signal processing (1978) 41–60.

[57] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, C. Toader, Customer characterization options for improving the tariff offer, IEEE Transactions on Power Systems 18 (1) (2003) 381–387.

[58] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve., Radiology 143 (4) (1982) 29–36. doi:10.1148/radiology.143.1.7063747.

Figures

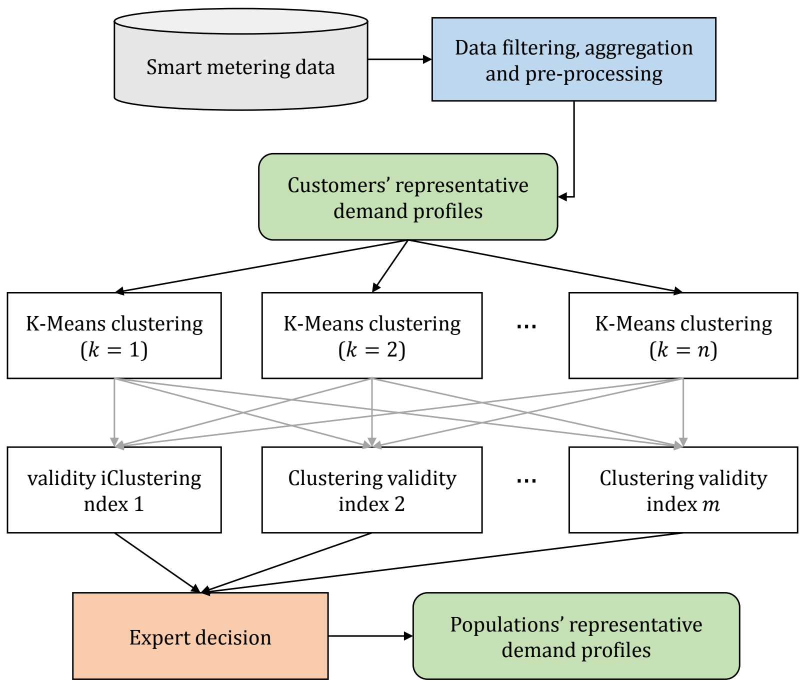
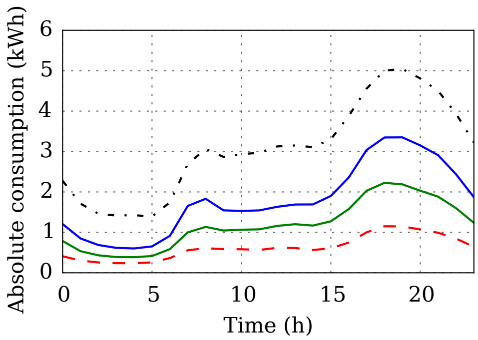
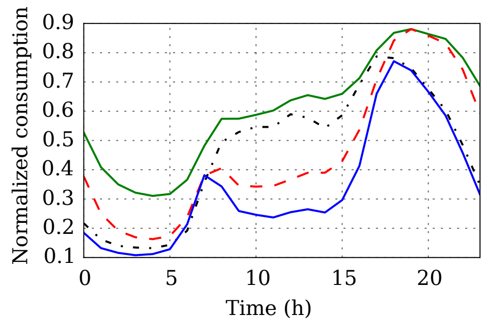


Figure 1: Generation of populations representative consumption profiles.



(a) Absolute consumption



(b) Customer normalized consumption

Figure 2: Example of populations representative consumption profiles using absolute and customer normalized consumption (resulting cluster centroids).

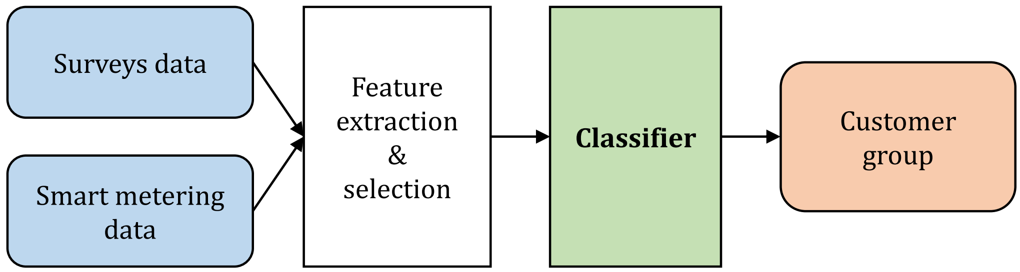


Figure 3: Electricity customer classifier.

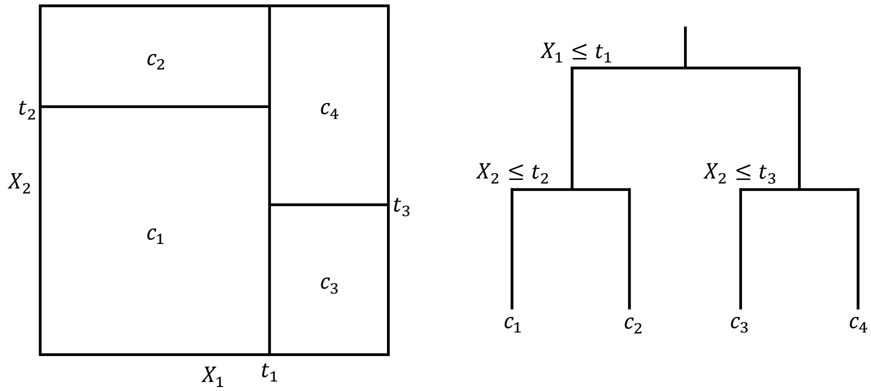


Figure 4: Left: data partitioned in four categories by binary splitting. Right: CART tree corresponding to the partition.

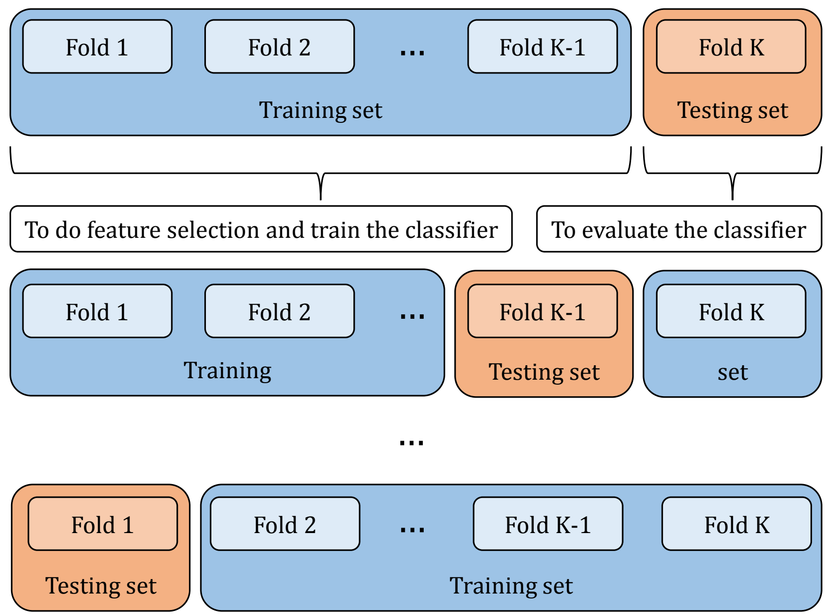


Figure 5: K-fold cross-validation.

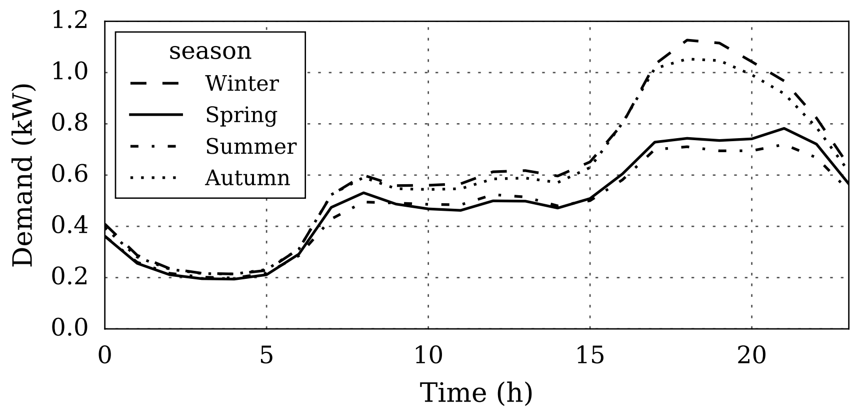
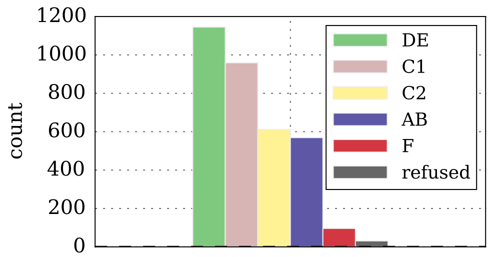
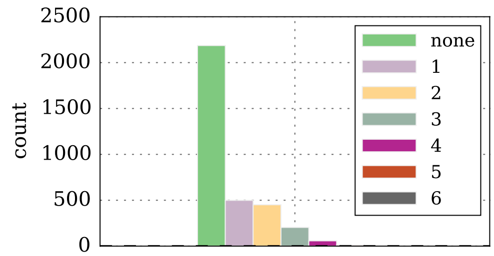


Figure 6: Hourly aggregated mean seasonal consumption of all customers.



(a) Social class



(b) Number of children

Figure 7: Distribution of the households for two survey responses.

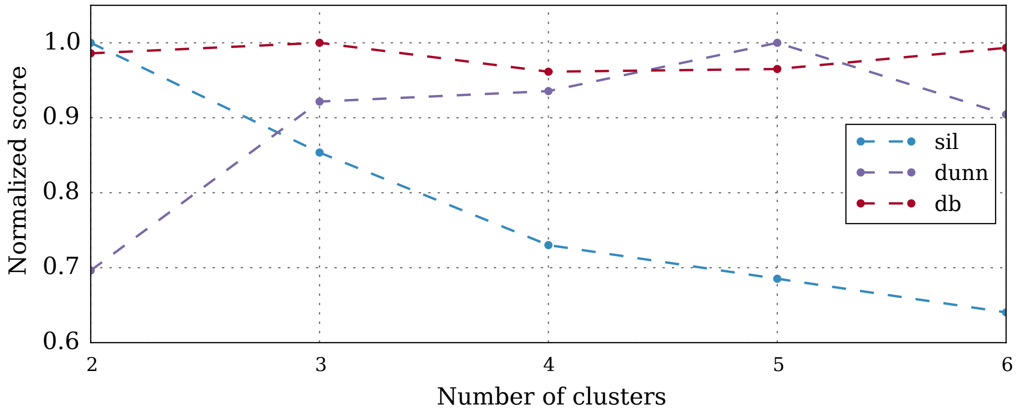


Figure 8: CVI for different number of clusters for the Winter consumption profiles.

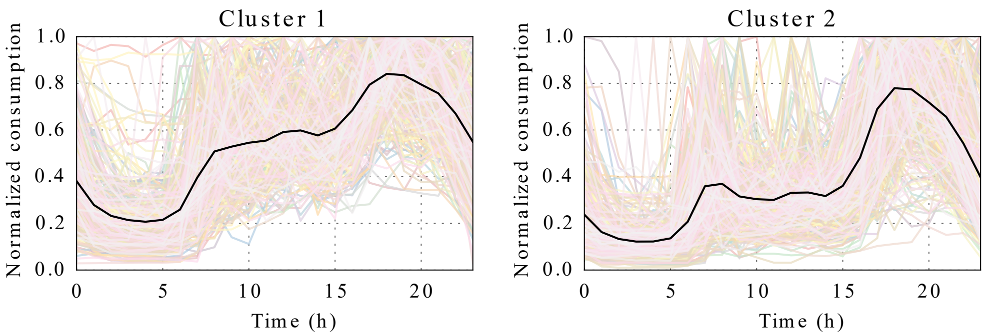


Figure 9: Winter clustering results with two clusters.

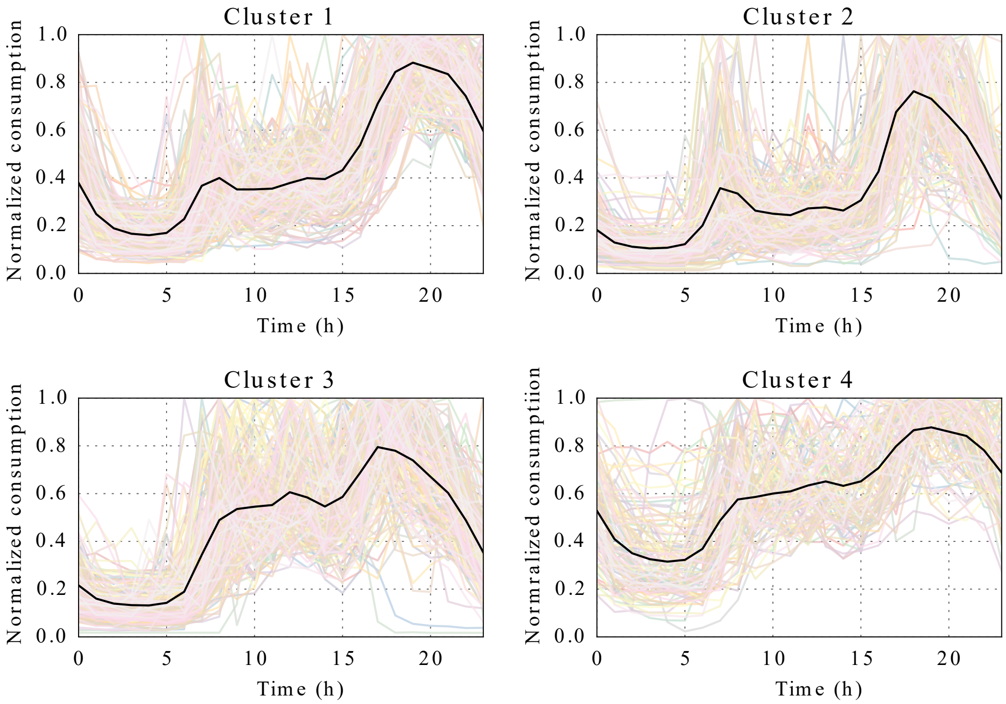


Figure 10: Winter clustering results with four clusters.

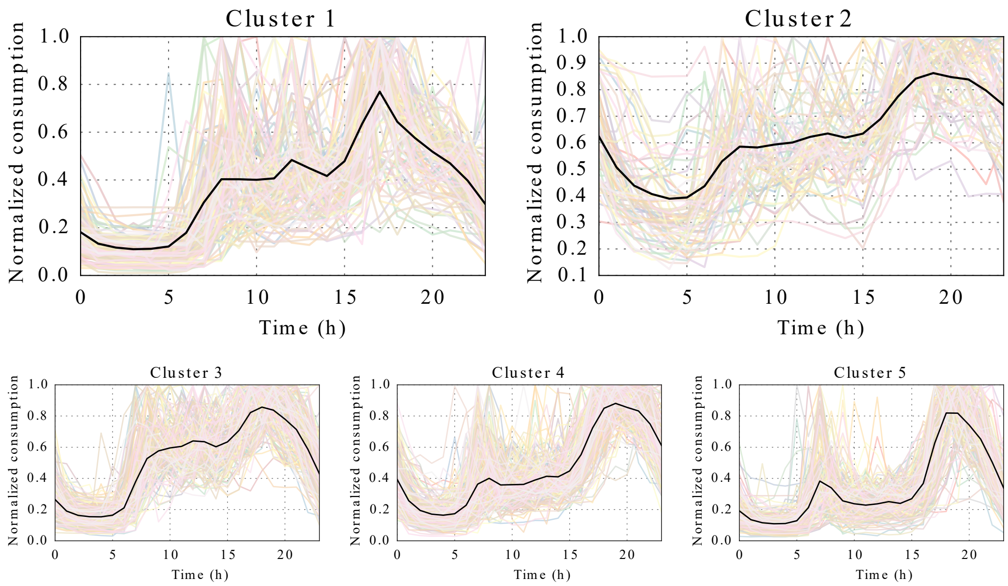
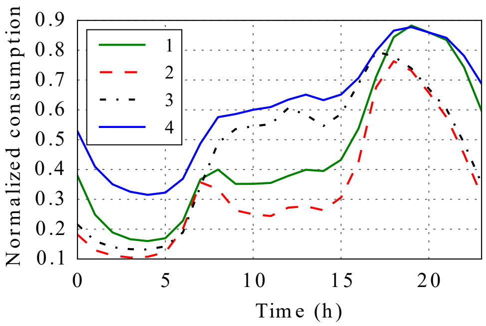
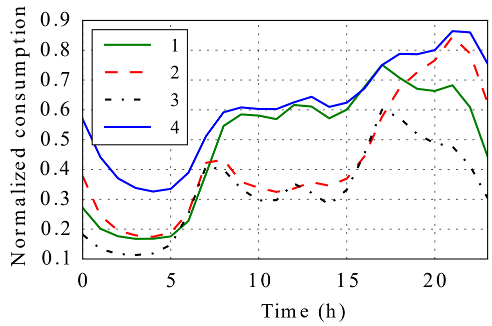


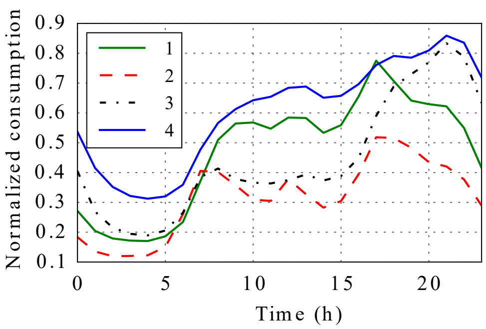
Figure 11: Winter clustering results with five clusters.



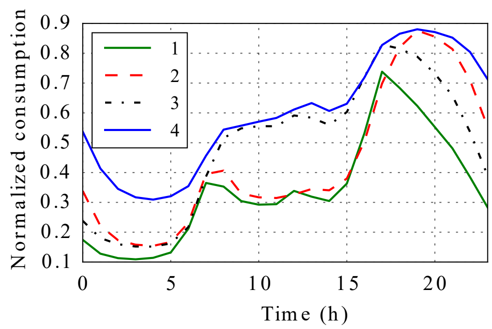
(a) Winter



(b) Spring

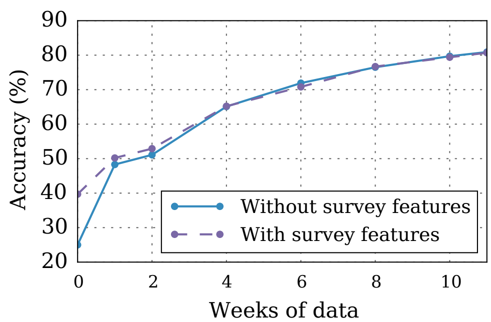


(c) Summer

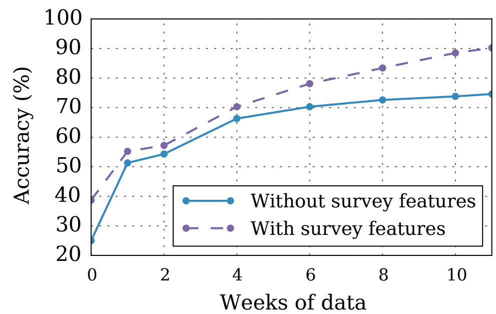


(d) Autumn

Figure 12: Populations representative consumption profiles.



(a) Metering data features: LI



(b) Metering data features: LP

Figure 13: LR classifier accuracy using filter FS with and without the survey features for Winter profiles.

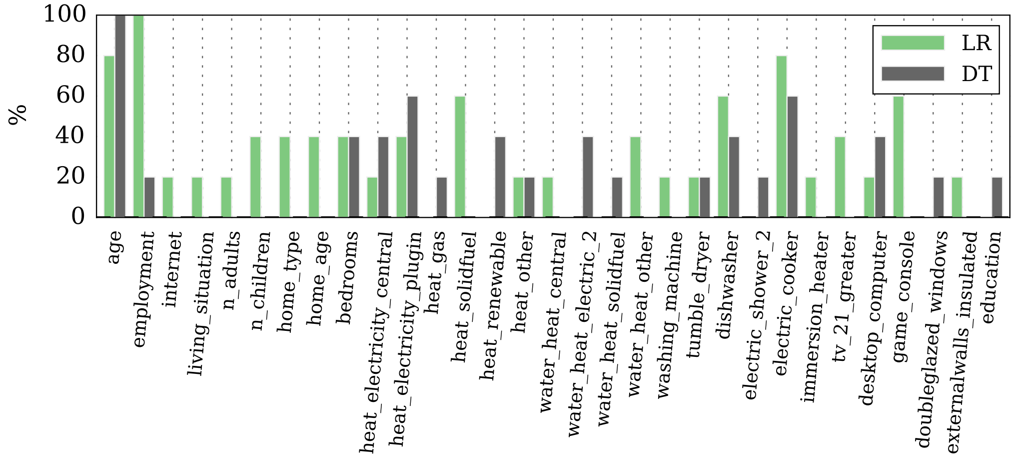


Figure 14: Forward FS for Winter with no metering data: rate of selection of features throughout the cross-validation process.

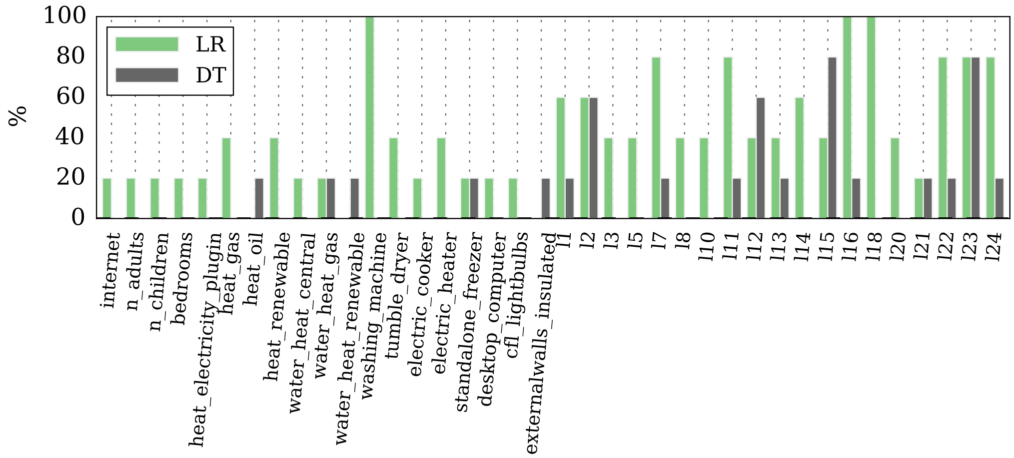


Figure 15: Forward FS for Summer with 4 weeks metering data (LP): rate of selection of features throughout the cross-validation process.

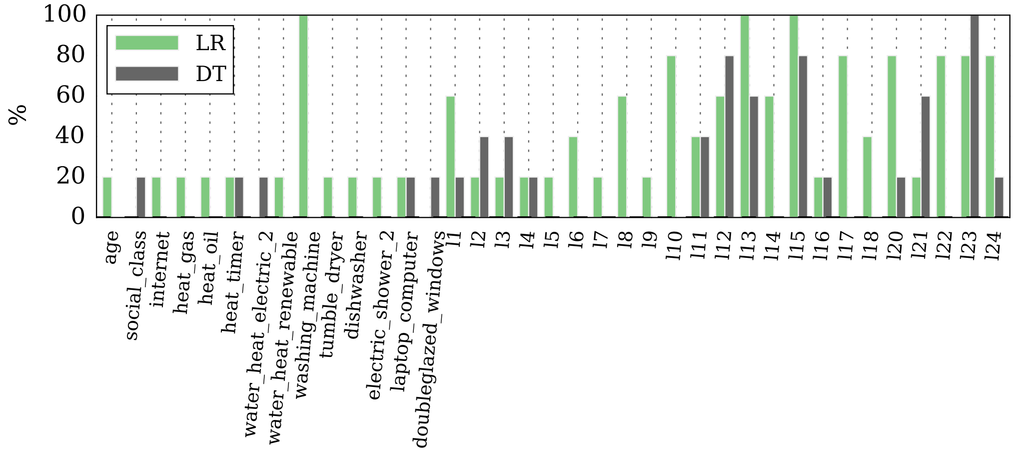


Figure 16: Forward FS for Autumn with 8 weeks metering data (LP): rate of selection of features throughout the cross-validation process.

Tables

Nomenclature			
Acronyms			
AMI	Advanced metering infrastructure	r	number of smart metering data features
EU	European Union	t	number of survey features
DSM	Demand side management	X	feature dataset of all customers
CVI	Clustering validity index	μ_i	ith consumption profile of the population
LP	Load profile		
LI	Load indexes	$\mathbf{S}$	set of the groups of customers
FS	Feature selection	S_i	ith clustered group of customers
LR	Logistic regression		
DT	Decision tree	J	number of clusters/customer groups
CER	Commission for Energy Regulation	$d_e(\mathbf{v}_1, \mathbf{v}_1)$	euclidean distance
ISSDA	Irish Social Science Data Archive	$D(\mathbf{S})$	Dunn index
		$DB(S)$	Davis Bouldin index
Symbols		$Sil(\mathbf{S})$	Silhouette index
$\mathbf{x}_i$	feature vector of customer i	y	categorical variable representing a group
$\mathbf{x}_i^m$	customer i smart metering data features	$i_1, i_2, \ldots, i_5$	load indices
$\mathbf{x}_i^s$	customer i surveys features	$P_{max/min/av}$	maximum, minimum and average consumption
N	number of customers		
p	dimension of feature vector	$l_1, l_2, \ldots, l_{24}$	load profile

Table 1: Survey features I: respondent

Feature	Description: {responses}
sex	Sex of respondent: {male, female}
age	Age of respondent in years: {18-25, 26-35, 36-45, 46-55, 56-65, 65 or more, refused}
employment	employment status of respondent: {Employee, self-employed, unemployed}
social_class	Social class of respondent: {AB, C1, C2, DE, F, refused}
education	Education level of respondent: {none, primary, secondary to intermediate cert junior cert level, secondary to leaving cert level, third level, refused}
income	Income of respondent before tax in euro: {0-15k, 15k-30k, 30k-50k, 50k-75k, 75k or more, refused}

Table 2: Survey features II: household

Feature	Description: {responses}
home_type	Household type: {apartment, semi-detached, detached, terraced, bungalow}
home_age	Household age in years: {0-4, 5-9, 10-29, 30-74, 75 or more}
bedrooms	Number of bedrooms : {1, 2, 3, 4, 5 or more, refused}
clf_lighbulbs	Fraction of CLF light bulbs: {none, about a quarter, about half, about three quarters}
doublegazed_windows	Fraction of doubleglazed windows: {none, about a quarter, about half, about three quarters}
attic_insulated	Presence and age of attic insulation: {yes (last 5 years), yes, no, don't know}
externalwalls_insuled	Presence and age of insulation of external walls: {yes, no, don't know}
internet	Internet connection in the household: {yes, no}

Table 3: Survey features III: heating

Feature	Description: {responses}
heat_electricity_central	Central electric heating : {yes, no}
heat_gas	Gas heating : {yes, no}
heat_oil	Oil heating : {yes, no}
heat_solidfuel	Solid fuel heating : {yes, no}
heat_renewable	Renewable energy heating : {yes, no}
heat_other	Other type of heating : {yes, no}
heat_timer	Use of heating timer : {yes, no}
water_heat_central	Central water heating : {yes, no}
water_heat_electric	Electric water heating: {yes, no}
water_heat_gas	Gas water heating: {yes, no}
water_heat_oil	Oil water heating: {yes, no}
water_heat_solidfuel	Solid fuel water heating: {yes, no}
water_heat_renewable	Renewable water heating: {yes, no}
water_heat_other	Other water heating source : {yes, no}

Table 4: Survey features IV: appliances

Feature	Description: {responses}
`washing_machine`	Number of washing machines : {0, 1, 2, 3 or more}
`tumble_dryer`	Number of tumble dryers : {0, 1, 2, 3 or more}
`dishwasher`	Number of dishwashers : {0, 1, 2, 3 or more}
`electric_shower`	Number of electric showers : {0, 1, 2, 3 or more}
`electric_cooker`	Number of electric cookers : {0, 1, 2, 3 or more}
`electric_heater`	Number of electric heaters : {0, 1, 2, 3 or more}
`standalone_freezer`	Number of standalone freezers : {0, 1, 2, 3 or more}
`water_pump`	Number of water pumps : {0, 1, 2, 3 or more}
`immersion_heater`	Number of immersion heaters : {0, 1, 2, 3 or more}
`tv_21_less`	Numbers of TVs with 21 or less inches: {0, 1, 2, 3, 4 or more}
`tv_21_greater`	Number of TVs with more than 21 inches: {0, 1, 2, 3, 4 or more}
`desktop_computer`	Number of desktop computers: {0, 1, 2, 3, 4 or more}
`laptop_computer`	Number of laptop computers: {0, 1, 2, 3, 4 or more}
`game_console`	Number of game consoles: {0, 1, 2, 3, 4 or more}

Table 5: Distribution of customers between the different clusters for the four seasons

Cluster	Winter	Spring	Summer	Autumn
1	30.93%	26.25%	26.14%	20.34%
2	25.50%	31.89%	18.83%	31.47%
3	28.17%	21.53%	27.17%	29.19%
4	15.39%	20.33%	27.86%	18.99%

Table 6: Normalized indices to characterize electricity customers' behaviour

Parameter	Definition	Periods
Daily P_{av}/P_{max}	$i_1 = P_{av,day}/P_{max,day}$	1 day
Daily $P_{min,day}/P_{max,day}$	$i_2 = P_{min,day}/P_{max,day}$	1 day
Night impact	$i_3 = 1/3 P_{av,night}/P_{av,day}$	1 day and 8 h night (from 23h to 06h)
Lunch impact	$i_4 = 1/8 P_{av,lunch}/P_{av,day}$	1 day and 3 h lunch from (12h to 15h)
Daily P_{min}/P_{av}	$i_5 = P_{min,day}/P_{av,day}$	1 day

Table 7: Smart metering data features used for classification

Smart metering data features		
Load indices (LI)	Normalized indices to characterize electricity constumers' behaviour.	i_1, i_2, i_3, i_4, i_5
Load profile (LP)	Normalized mean hourly aggregated consumption.	$l_1, l_2, \ldots, l_{24}$

Table 8: Mean 10-fold cross-validation accuracy of classifiers using load indices as metering data features (number of selected features)

Smart metering data features: Load indices					
W	Model	Winter	Spring	Summer	Autumn
No FS					
0	LR	39.2±0.8 (47)	37.6±1.1 (47)	37.0±1.9 (47)	38.5±0.7 (47)
	DT	36.0±1.5 (47)	34.7±1.1 (47)	33.8±2.4 (47)	36.7±1.7 (47)
1	LR	45.3±6.9 (52)	54.9±1.1 (52)	53.1±1.5 (52)	53.4±1.3 (52)
	DT	46.5±1.7 (52)	53.3±1.6 (52)	51.8±1.5 (52)	51.4±2.0 (52)
4	LR	64.9±1.8 (52)	64.6±1.3 (52)	65.7±2.1 (52)	62.0±1.4 (52)
	DT	62.8±0.8 (52)	62.4±2.7 (52)	63.3±2.7 (52)	59.3±1.9 (52)
8	LR	75.8±1.3 (52)	71.8±0.8 (52)	71.0±0.9 (52)	57.1±19.0 (52)
	DT	73.3±1.3 (52)	70.4±0.5 (52)	69.3±2.3 (52)	67.7±1.6 (52)
10	LR	78.3±1.4 (52)	75.4±0.8 (52)	64.9±19.1 (52)	73.4±1.8 (52)
	DT	75.1±1.0 (52)	72.9±0.8 (52)	71.7±1.3 (52)	72.1±1.8 (52)
Filter FS					
0	LR	38.6±1.8 (17)	36.2±2.3 (18)	35.9±1.1 (17)	34.6±7.6 (23)
	DT	36.7±1.7 (17)	35.7±0.9 (18)	34.1±2.2 (17)	35.1±0.5 (23)
1	LR	49.8±0.8 (21)	56.5±1.7 (20)	53.9±2.3 (21)	53.4±1.3 (19)
	DT	46.3±2.7 (21)	52.8±1.4 (20)	51.0±0.3 (21)	50.6±1.4 (19)
4	LR	58.4±12.4 (26)	65.9±1.3 (18)	66.8±0.3 (16)	62.6±1.7 (19)
	DT	62.1±2.7 (26)	62.0±0.8 (18)	64.2±1.7 (16)	60.1±1.0 (19)
8	LR	76.5±2.4 (19)	72.7±1.8 (17)	72.0±0.5 (17)	59.4±19.7 (26)
	DT	73.8±0.9 (19)	69.5±2.0 (17)	69.8±1.7 (17)	67.7±1.2 (26)
10	LR	79.1±1.5 (17)	76.0±2.0 (19)	75.2±0.7 (15)	74.3±1.6 (22)
	DT	75.3±1.9 (17)	72.6±1.1 (19)	71.9±1.8 (15)	72.1±1.2 (22)
Forward FS					
0	LR	38.2±1.1 (9)	36.7±1.3 (5)	34.8±1.5 (10)	37.7±4.3 (5)
	DT	36.6±1.1 (6)	35.8±0.5 (4)	32.9±1.1 (5)	37.4±4.3 (2)
1	LR	49.5±1.1 (11)	56.0±2.4 (9)	54.3±1.7 (10)	53.0±3.6 (10)
	DT	46.9±1.5 (6)	52.6±0.9 (5)	52.3±0.9 (5)	50.6±2.2 (4)
4	LR	50.7±16.7 (6)	65.5±1.3 (11)	66.3±1.5 (7)	62.5±1.4 (7)
	DT	61.7±1.9 (4)	62.9±2.1 (4)	63.1±1.2 (5)	60.4±0.8 (4)
8	LR	76.4±1.3 (8)	72.1±2.5 (8)	72.2±0.9 (9)	70.7±1.0 (8)
	DT	71.5±0.8 (4)	69.8±1.3 (4)	69.5±1.2 (4)	67.4±1.9 (4)
10	LR	79.2±1.8 (9)	75.6±1.5 (9)	76.0±1.4 (9)	74.3±1.5 (8)
	DT	75.8±1.3 (4)	72.7±2.6 (4)	72.3±1.2 (4)	71.8±0.8 (3)

Table 9: Mean 10-fold cross-validation accuracy of classifiers using the load profile as metering data features (mean number of selected features)

Smart metering data features: Load profile					
W	Model	Winter	Spring	Summer	Autumn
No FS					
0	LR	38.7±2.1 (47)	37.2±2.6 (47)	36.6±2.3 (47)	38.6±0.9 (47)
	DT	36.0±1.3 (47)	34.7±1.2 (47)	34.4±1.8 (47)	35.1±0.8 (47)
1	LR	53.9±0.9 (71)	60.8±2.4 (71)	58.7±2.3 (71)	60.6±1.3 (71)
	DT	48.0±2.6 (71)	54.8±2.2 (71)	52.0±0.9 (71)	53.3±2.0 (71)
4	LR	70.8±1.9 (71)	72.5±1.6 (71)	72.3±1.0 (71)	70.6±1.2 (71)
	DT	63.6±2.5 (71)	65.2±1.7 (71)	65.6±2.1 (71)	64.5±1.7 (71)
8	LR	83.4±1.5 (71)	80.8±1.3 (71)	79.4±1.1 (71)	78.4±1.2 (71)
	DT	74.1±1.6 (71)	72.7±1.5 (71)	70.6±1.8 (71)	71.7±1.9 (71)
10	LR	76.3±22.2 (71)	73.1±24.7 (71)	82.9±1.0 (71)	83.2±1.2 (71)
	DT	76.6±1.2 (71)	74.9±1.0 (71)	73.2±1.1 (71)	75.4±1.8 (71)
Filter FS					
0	LR	39.0±1.2 (16)	37.4±0.9 (17)	35.3±1.3 (18)	38.9±1.4 (16)
	DT	36.3±0.7 (16)	35.4±0.7 (17)	34.2±1.2 (18)	36.5±1.7 (16)
1	LR	53.9±0.8 (29)	60.8±0.6 (28)	59.7±1.1 (32)	60.9±1.5 (28)
	DT	49.0±0.9 (29)	55.0±1.5 (28)	52.6±1.2 (32)	53.6±2.6 (28)
4	LR	62.2±16.5 (40)	72.9±1.6 (29)	73.3±1.0 (30)	71.4±2.2 (29)
	DT	63.9±0.4 (40)	64.1±1.7 (29)	64.9±1.5 (30)	64.3±1.7 (29)
8	LR	83.1±2.4 (32)	81.6±0.8 (32)	79.6±1.8 (34)	78.9±1.9 (33)
	DT	73.4±2.3 (32)	72.8±0.9 (32)	70.8±1.0 (34)	71.4±1.9 (33)
10	LR	88.3±0.9 (37)	86.1±0.6 (42)	83.1±0.9 (37)	83.8±0.8 (38)
	DT	76.3±1.1 (37)	76.5±1.5 (42)	72.4±0.7 (37)	76.8±1.0 (38)
Forward FS					
0	LR	37.3±5.2 (9)	36.8±0.9 (8)	34.1±1.8 (8)	37.8±1.0 (7)
	DT	37.4±2.3 (4)	36.5±0.9 (3)	32.2±1.2 (4)	36.2±2.0 (3)
1	LR	50.9±1.4 (11)	59.1±1.6 (13)	56.4±2.4 (13)	57.9±1.6 (12)
	DT	48.2±1.9 (5)	52.6±1.9 (6)	51.9±2.4 (6)	51.6±1.3 (6)
4	LR	69.8±1.6 (16)	70.3±1.3 (11)	71.7±1.7 (16)	70.2±1.4 (13)
	DT	63.4±1.4 (7)	64.3±1.5 (6)	64.9±1.9 (5)	62.9±2.3 (6)
8	LR	83.4±1.1 (14)	80.8±1.1 (16)	77.9±2.1 (15)	77.8±1.2 (14)
	DT	73.2±1.4 (5)	71.9±2.1 (8)	70.4±2.5 (7)	70.8±1.9 (6)
10	LR	87.3±0.8 (16)	85.2±0.9 (17)	81.1±2.0 (16)	82.8±1.3 (15)
	DT	76.1±1.1 (6)	75.1±2.2 (6)	73.4±2.2 (5)	74.2±1.1 (7)

Table 10: Filter FS for Winter with no metering data: variables found to be significant for at least one of the classifiers of the MNLogit

Filter FS: Winter with no metering data		
age	employment	social_class
living_situation	n_children	bedrooms
water_heat_oil	dishwasher	electric_shower_1
electric_shower_2	electric_cooker	electric_heater
tv_21_greater	desktop_computer	game_console
cfl_lightbulbs	cfl_lightbulbs	cfl_lightbulbs

Table 11: Filter FS for Spring with 1 week metering data (LI): variables found to be significant for at least one of the classifiers of the MNLogit

Filter FS: Spring with one week metering data (LI)		
age	employment	living_situation
n_children	home_type	home_age
bedrooms	heat_solidfuel	water_heat_solidfuel
washing_machine	tumble_dryer	dishwasher
electric_shower_2	electric_cooker	tv_21_less
externalwalls_insulated	education	income
i1	i2	i3
i4	i4	i4

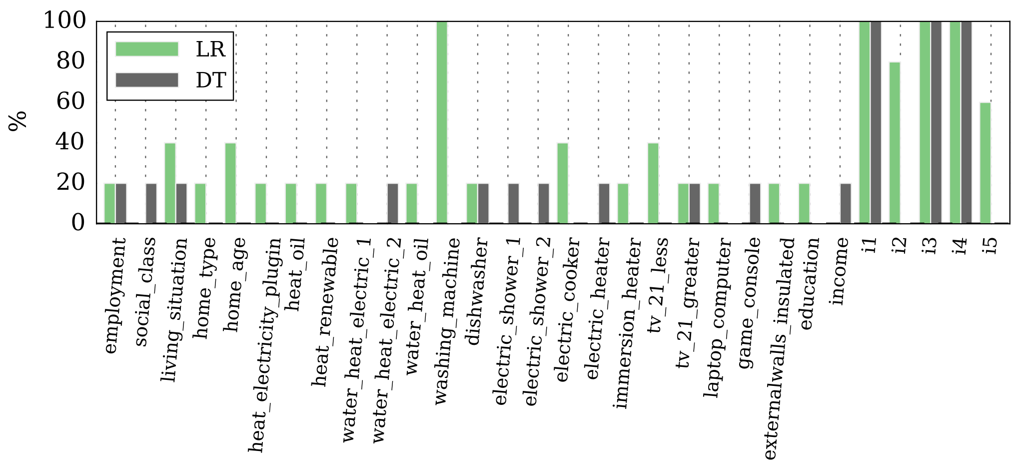


Figure 17: Forward FS for Spring with 1 week metering data (LI): rate of selection of features throughout the cross-validation process.

Table 12: Filter FS for Summer with four weeks metering data (LP): variables found to be significant for at least one of the classifiers of the MNLogit

Filter FS: Summer with four weeks metering data (LP)		
age	social_class	internet
living_situation	n_children	home_type
water_heat_electric_2	water_heat_oil	washing_machine
electric_cooker	standalone_freezer	l1
l3	l8	l9
l10	l11	l12
l13	l14	l15
l16	l17	l18
l19	l20	l21
l22	l23	l24

Table 13: Filter FS for Autumn with eight weeks metering data (LP): variables found to be significant for at least one of the classifiers of the MNLogit

Filter FS: Autumn with eight weeks metering data (LP)		
internet	living_situation	heat_timer
water_heat_electric_2	water_heat_gas	water_heat_oil
washing_machine	tumble_dryer	electric_cooker
game_console	cfl_lightbulbs	attic_insulated
externalwalls_insulated	education	l2
l5	l6	l9
l10	l11	l12
l13	l14	l15
l16	l17	l18
l19	l20	l21
l22	l23	l24