Revealing Household Characteristics
from Smart Meter Data

Christian Beckel[a,*], Leyna Sadamori[a], Thorsten Staake[b], Silvia Santini[c]

[a]*Institute for Pervasive Computing, Department of Computer Science, ETH Zurich, Universitätstr. 6, 8092 Zurich, Switzerland*
[b]*Energy Efficient Systems Group, University of Bamberg, An der Weberei 5, 96047 Bamberg, Germany*
[c]*Wireless Sensor Networks Lab, TU Darmstadt, Rundeturmstr. 10, 64283 Darmstadt, Germany*

Abstract

Utilities are currently deploying smart electricity meters in millions of households worldwide to collect fine-grained electricity consumption data. We present an approach to automatically analyzing this data to enable personalized and scalable energy efficiency programs for private households. In particular, we develop and evaluate a system that uses supervised machine learning techniques to automatically estimate specific "characteristics" of a household from its electricity consumption. The characteristics are related to a household's socio-economic status, its dwelling, or its appliance stock. We evaluate our approach by analyzing smart meter data collected from 4,232 households in Ireland at a 30-minute granularity over a period of 1.5 years. Our analysis shows that revealing characteristics from smart meter data is feasible, as our method achieves an accuracy of more than 70% over all households for many of the characteristics and even exceeds 80% for some of the characteristics. The findings are applicable to all smart metering systems without making changes to the measurement infrastructure. The inferred knowledge paves the way for targeted energy efficiency programs and other services that benefit from improved customer insights. On the basis of these promising results, the paper discusses the potential for utilities as well as policy and privacy implications.

Keywords: Data-driven energy efficiency, Domestic electricity consumption, Electricity load profiles, Automated customer segmentation, Supervised machine learning

*Corresponding author, Phone: +41 44 632 7871, Fax: +41 44 632 1659

Email addresses: beckel@inf.ethz.ch (Christian Beckel), sadamori@inf.ethz.ch (Leyna Sadamori), thorsten.staake@uni-bamberg.de (Thorsten Staake), santinis@wsn.tu-darmstadt.de (Silvia Santini)

1. Introduction

Customer insights help utilities to optimize their energy efficiency programs in many ways [1, 2]. With knowledge of the socio-economic characteristics of individual households, for instance, utilities can automatically tailor savings advice to specific addressees (e.g., to families with children, or to retirees). Further, they can offer consumption feedback that includes references to similar households or consider the financial reach of their customers when suggesting improvements in the appliance stock. Many studies have shown that such specific approaches improve the performance of efficiency campaigns [3, 4, 5]. Yet, such targeted measures require detailed information on individual customers, which might be gathered for research studies and local saving campaigns, but which is often not available for large-scale, cost sensitive efficiency programs that are directed to millions of households.

In fact, utilities' knowledge about their customers is often limited to their address and billing information. This is particularly true in Europe, where open information repositories like public tax registers do not exist or cannot be easily accessed. On the other hand, conducting surveys to acquire customer information is typically time-consuming and expensive, and often only a small fraction of customers participate [6]. We argue that utilities can instead utilize the electricity consumption data of a household to reveal customer information that is relevant to optimizing their energy efficiency programs. This is valuable for utilities, because they are already deploying millions of smart electricity meters in private households along with infrastructure to collect, process, and store their electricity consumption data. [7, 8, 9]. Currently, utilities use this data mainly to improve their meter-to-cash processes, to enable advanced tariff schemes, and to provide customers with detailed information on their electricity consumption. Analyzing smart meter data that is collected anyway can therefore contribute to the value of the metering infrastructure without requiring any changes to the smart meters that have already been deployed.

In this paper, we develop and evaluate a system to automatically infer household characteristics from smart meter data. Examples of such characteristics include the household's socio-economic status, its dwelling properties, and information on the appliance stock. Our analysis takes as input the electricity consumption of a household and estimates the value of several characteristics of interest. Depending on the characteristic, this value is either the class to which the household most likely belongs to (e.g., employment status) or a numerical value (e.g., the number of persons living in the household). To infer the value of household characteristics from consumption data, we extract features from the data itself and pass them as input to a classifier or regression model. An example of such a feature is the average consumption of a household between 10 a.m. and 2 p.m. divided by its daily average consumption. This particular feature helps to reveal household occupancy during lunch time and thus contributes to the estimation of characteristics such as the employment status of the inhabitants. We investigate 18 different characteristics which we have selected because they are relevant to utilities [10]. We have evaluated our

system according to these characteristics using smart meter data available at a 30-minute granularity from 4,232 Irish households over a period of 1.5 years. This data set is publicly available and has been collected in the context of a smart metering trial conducted by the Irish Commission for Energy Regulation (CER)[1]. In the following, we refer to this data set as the *CER data set*. Along with smart meter data, the data set contains information on the characteristics of each household collected through questionnaires before and after the study. This information is crucial for our work, because it represents ground truth data we can use to validate our findings.

The contribution of this paper is a comprehensive system for automatically revealing household characteristics from smart meter data and an elaborate evaluation of our approach. In our previous work [11], we presented a preliminary study to demonstrate the feasibility of revealing household properties from smart meter data. In this paper, we improve upon our previous work in multiple respects: First, we present new components of our system. We extend the feature set, replace the feature selection method, and add a classifier. Second, we perform a detailed analysis to evaluate the applicability of our results. In particular, we advocate and discuss new performance measures (e.g., to handle imbalanced classes), we investigate six additional characteristics that are of interest to utilities, and we propose and evaluate the utilization of the classifier confidence to identify small groups of customers with improved performance. We also propose a regression model in order to estimate characteristics with continuous values (e.g., the number of persons in a household). Finally, we show the stability of the results over all 75 weeks included in the data set, and we show significant performance gains that can be achieved when performing the analysis on the whole measurement period instead of a single week of data only, as it was done in [11].

The results provided in this paper show that revealing household characteristics from smart meter data is feasible with sufficient accuracy. This holds in particular for characteristics related to the number of persons living in a household and for characteristics related to the occupancy of the household (which also includes information on the employment status of the chief income earner). We show that it is possible to infer 8 of the 18 characteristics with an accuracy between 72% and 82%. Overall, our approach performs roughly 30 percentage points better than assigning characteristics to the households at random. Some applications require identifying households that feature a specific characteristic with high accuracy. This is for instance necessary when a group of households (e.g., those inhabited by a single person) are the target of a marketing campaign. Here, reducing the number of false positives (i.e., of the cases in which a household is erroneously estimated to belong to the target class) is crucial. We show that by exploiting the confidence of the estimation obtained from the classifiers, it is possible to reduce the number of false positives significantly.

According to the results reported in this paper, utilities can reliably esti-

---

[1] www.ucd.ie/issda/data/commissionforenergyregulationcer/

3

mate household characteristics from smart meter data. Thus, they will be able to improve their energy efficiency campaigns and make them applicable to the mass market as they scale to thousands or millions of customers with little additional effort. Ultimately, creating these services to help their customers use energy more efficiently is crucial for utilities' attempts to comply with regulatory targets [12]. In addition, the system provided in this paper allows utilities to improve customer retention, which is becoming more relevant in a liberalized energy market [13]. To the best of our knowledge, this is the first study that provides a quantitative analysis of the possibility of revealing household characteristics from electricity consumption data on such a large data set and at such a high accuracy.

The remainder of this paper is structured as follows. Section 2 reviews related work. We then present the data set we use in our study in section 3 and our methodology in section 4. Next, we describe our evaluation setup and performance measures in section 5. Section 6 presents the results of the analysis followed by a discussion of the results in section 7. Finally, section 8 concludes the paper and gives an outlook on future work.

2. Related Work

Over the past years, an increasing number of researchers have applied machine learning and data mining techniques to model and analyze residential electricity consumption data. This has been made possible thanks to the increasing availability of electricity consumption data. A popular line of research in this context is one focusing on *non-intrusive load monitoring* (NILM). Using aggregated electricity consumption data of individual households (e.g., measured at 1 reading per second or millisecond), researchers have tackled the problem of disaggregating the consumption of individual appliances. This information allows in turn to provide detailed consumption feedback to the households [14, 15, 16]. The work we present in this paper is considerably different from NILM, because we aim to infer high-level household characteristics from the electricity consumption instead of disaggregating it into its individual end use.

Other authors have focused on the analysis of coarse-grained consumption data (i.e., data sampled at a granularity of several minutes or higher). Here, we distinguish between (1) analyzing consumption data only and (2) relating it to side-information such as the geographic location of the dwelling or the socio-economic status of the household. Since the first approach imposes less requirements on the collected data, many authors have investigated unsupervised techniques such as clustering to detect patterns and usage categories in the consumption profiles [17, 18, 19, 20]. Chicco, for instance, provides an overview of clustering techniques used to group residential or commercial customers according to their electricity consumption pattern [19]. Grouping consumers by their load profile enables utilities to formulate tariffs for specific customer categories, check the effect of tariff modifications, and ultimately optimize their supply management. Using similar techniques, both Kwac et al. [18] and Cao

4

et al. [17] have focused identifying the "right" customers for demand-side management campaigns. Whereas Kwac et al. aim at detecting stable profiles over a certain time period, Cao et al. focus on identifying households with a similar time of peak usage. Finally, De Silva et al. aim at predicting future electricity usage of private households using a data mining framework and an incremental learning algorithm [20]. In contrast to all these approaches, our work goes beyond detecting consumption patterns or usage categories. We utilize such patterns to estimate specific characteristics of the socio-economic status, the dwelling, or appliance stock of the households.

In recent studies, researchers increasingly investigated the combination of electricity consumption with side-information [11, 21, 22, 23, 24, 25, 26, 27]. Sanchez et al. add information about the households gathered through questionnaires to features they derive from electricity consumption data [22]. They then cluster 625 Spanish households using a well-known technique called self-organizing maps (SOMs) [28]. Räsänen et al. also use SOMs to cluster households [23]. However, for input the authors rely on dwelling characteristics only, with the goal of providing personalized electricity use information to households within the same cluster. Kolter et al. apply a regression model to estimate monthly consumption data from household characteristics derived from public databases in the United States [24]. Comparing this estimation with the actual consumption of a household enables personalized feedback to be provided to the inhabitants of the household. Relying on a similar regression model, Kavousian et al. analyze the effect of so-called determinants on the household electricity consumption [27]. In particular, the authors define four major categories of determinants that affect the overall consumption: (1) Weather and location, (2) dwelling characteristics, (3) appliance and electronics stock, and (4) occupancy and behavior. After applying their model on 1,628 households in the United States, the authors come to the conclusion that weather and dwelling characteristics have a larger influence on residential electricity consumption compared to the appliance stock and occupancy behavior. It is important to note, however, that the data used in their study also accounts for electricity consumed by heating and cooling, which represent a large portion of the overall electricity consumption.

McLoughlin et al. also investigated the correlation between electricity consumption data and household characteristics [25]. Like Kavousian et al., the authors used a multiple linear regression analysis to model the electricity consumption of households on the basis of their characteristics. Relying on the same data as the present study does – which does not account for thermal loads – the authors found a strong relationship between four electricity consumption parameters (total consumption, maximum demand, load factor, and time of use) and different dwelling, household, and appliance stock characteristics. In his dissertation, McLoughlin further investigated methods to automatically cluster households in order to segment them into profile groups according to their electricity consumption [26]. McLoughlin then investigates the distribution of household characteristics over the clusters with the goal of characterizing electricity use depending on the customer characteristics. In contrast to both

5

Kavousian et al. and McLoughlin et al., we propose a method that utilizes the correlation between electricity consumption data and household characteristics to estimate the characteristics from the consumption data. Albert et al. recently presented an approach that has similar goals as ours [21]. The authors first remove the impact of weather on a household's electricity consumption using a linear regression model. On the residuals they utilize a Hidden Markov Model to infer specific occupancy states per household. All parameters gained from this analysis then serve as input to an AdaBoost classifier in order to estimate specific household characteristics. To evaluate their work, the authors rely on the same data set as Kavousian et al., which consists of smart meter data and household characteristics of 950 Google employees. In contrast to Albert et al., our system relies on a different set of features and it integrates a feature selection method as well as multiple classifiers in addition to the AdaBoost classifier. Furthermore, we use more performance measures and a data set that is much larger than the Google data set to evaluate our work.

The approaches presented above either include household characteristics as a part of a regression model or they rely on a relatively small set of households. In our study we present a system that relies on supervised machine learning techniques to estimate household characteristics from electricity consumption data. We further utilize consumption data and household characteristics of 4,232 households to train our classifiers and evaluate our approach.

3. The CER data set

Our study relies on the CER data set, which was collected during a smart metering trial conducted in Ireland by the Irish Commission for Energy Regulation (CER). It contains measurements of electricity consumption gathered from 4,232 households every 30 minutes between July 2009 and December 2010 (75 weeks in total). The purpose of the study was to investigate the effect of consumption feedback on household electricity consumption. Each participating household was asked to fill out a questionnaire before and after the study. The questionnaire contained questions about the household's socio-economic status, appliance stock, properties of the dwelling, and the consumption behavior of the occupants.

In contrast to other studies that investigated large-scale electricity consumption data (such as [21, 24, 27]), the CER data set to the best of our knowledge does not account for energy that is consumed by heating and cooling systems. The heating systems of the participating households either use oil or gas as their source of energy or their consumption is measured by a separate electricity meter. The households involved in the study were reported to have no cooling system installed.

4. System design

Our analysis relies on supervised machine learning techniques to infer a household's characteristics from its electricity consumption data. Figure 1 de-
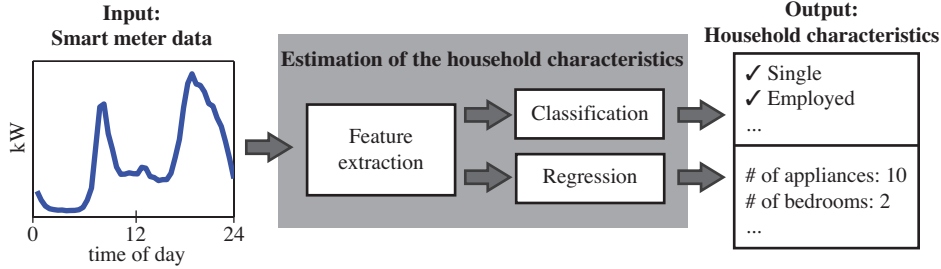
6

Figure 1: Overview of the household characteristic estimation presented in this paper.

picts the household characteristic estimation process. First, we compute a set of *features* on the electricity consumption records of a household. This is a typical step performed in supervised machine learning to obtain a set of discriminative values for each sample (i.e., household). The features then serve as input to a *classifier* or *regression model*, depending on the characteristic. As output, our system provides an estimate of the class or the value of each characteristic.

4.1. Features

Table 1 lists the features we compute on the electricity consumption data. We divide the features into five groups: consumption figures (10 features), ratios of consumption figures (7 features), features related to temporal dynamics (4 features), statistical properties (3 features), and the first ten principal components [29]. Our system assumes the data to be available at a granularity of one measurement every 30 minutes and it computes each feature on one week of data. However, it can be easily adapted to cope with other data granularities and time periods. Part of the features have been used in previous work on the analysis of electricity consumption data [10, 11, 22].

Many statistical methods assume the input data to follow a normal distribution [30]. For this reason, researchers often apply a non-linear transformation (e.g., a logarithmic or square root transformation) to each of the features if it improves normality [30]. To find the right transformation, we (visually) compare the distribution of the transformed feature with the normal distribution using a normal quantile plot [31]. Figure 2 shows the normal quantile plot for features c_total and r_morning/noon transformed by a logarithmic and a square root transformation, respectively. The linearity of the sample quantiles of the features (x-axis) versus the theoretical quantiles of a normal distribution (y-axis) implies that the transformed features are (roughly) normally distributed. After the transformation, we normalize each feature such that it has zero mean and unit variance. Data normalization is required by some of the classifiers we consider in our study, for example when their objective function calculates a distance between two samples based on their features.

7

Table 1: List of features that form the input vectors of the classifiers. $\bar{P}$ denotes the 30-minute mean power samples provided by the data set. Where not otherwise stated, the feature is computed over the weekdays only. The last column shows if a logarithmic (log) or square root (sqrt) transformation has been applied to the feature.

| Description	Name	Transformation
(1) Consumption figures		
$\bar{P}$ (daily, week)	`c_total`	sqrt(x)
$\bar{P}$ (daily, weekdays)	`c_weekday`	sqrt(x)
$\bar{P}$ (daily, weekend)	`c_weekend`	sqrt(x)
$\bar{P}$ for (6 a.m. $-$ 10 p.m.)	`c_day`	sqrt(x)
$\bar{P}$ for (6 p.m. $-$ 10 p.m.)	`c_evening`	sqrt(x)
$\bar{P}$ for (6 a.m. $-$ 10 a.m.)	`c_morning`	sqrt(x)
$\bar{P}$ for (1 a.m. $-$ 5 a.m.)	`c_night`	log(x)
$\bar{P}$ for (10 a.m. $-$ 2 p.m.)	`c_noon`	sqrt(x)
Maximum of $\bar{P}$, week	`c_max`	x
Minimum of $\bar{P}$, week	`c_min`	log(x)
(2) Ratios		
Mean $\bar{P}$ over maximum $\bar{P}$	`r_mean/max`	log(x)
Minimum $\bar{P}$ over mean $\bar{P}$	`r_min/mean`	sqrt(sqrt(x))
`c_morning` / `c_noon`	`r_morning/noon`	log(x)
`c_evening` / `c_noon`	`r_evening/noon`	log(x)
`c_noon` / `c_total`	`r_noon/day`	sqrt(x)
`c_night` / `c_day`	`r_night/day`	log(x)
`c_weekday` / `c_weekend`	`r_weekday/weekend`	log(x)
(3) Temporal properties		
Proportion of time with $\bar{P} > 0.5$kW	`t_above_0.5kw`	x
Proportion of time with $\bar{P} > 1$kW	`t_above_1kw`	x
Proportion of time with $\bar{P} > 2$kW	`t_above_2kw`	x
Proportion of time with $\bar{P} >$ mean	`t_above_mean`	x
(4) Statistical properties		
Variance	`s_variance`	sqrt(sqrt(x))
$\sum(|\bar{P}_t - \bar{P}_{t-1}|)$ for all t	`s_diff`	sqrt(x)
Cross-correlation of subsequent days	`s_x-corr`	x
$\#\bar{P}$ with $(\bar{P}_t - \bar{P}_{t\pm1} > 0.2$ kW)	`s_num_peaks`	x
(5) Principal components		
First 10 principal components	`pca_i` *(i = 1..10)*	x
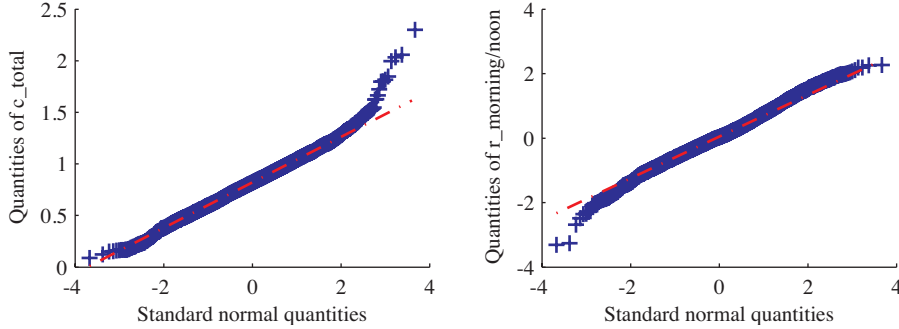
Figure 2: Normal quantile plots showing that features `c_total` (left) and `r_morning/noon` (right) are (roughly) normally distributed after applying the log and square root transformations, respectively.

4.2. Household characteristics and class labels

A classifier estimates a characteristic of a household by assigning the household to a specific *class* out of a set of classes. Table 2 shows the 18 characteristics we evaluate in this study along with the corresponding classes and class definitions for each characteristic. The characteristics capture socio-economic status of the household (e.g., `age_person`, `employment`), dwelling properties (e.g., `#bedrooms`, `floor_area`), or characteristics related to the behavior or appliance stock (e.g., `#appliances`, `unoccupied`). `#adults` and `#children` represent the number of adults and children in the household, respectively. The table also shows the number of samples for each class, where each sample corresponds to one household in the CER data set.

In a previous study, we identified the characteristics that are interesting for utilities by conducting interviews with four energy consultants [10]. The interviews revealed, for instance, that knowing the composition of a household (e.g., `single`, `family`) is particularly relevant to energy consultants, because families are potentially more interested than singles in receiving information about energy consulting services. Furthermore, we selected characteristics with well-separable classes, which means that the samples from different classes have (on average) a high distance in the feature space. As an example, figure 3 illustrates class separability of the characteristic `single` for features `c_total` and `r_evening_noon` based on the empirical cumulative distribution (ECD) for each of the two features. The left plot shows that the ECD of the first class (*Single*) significantly differs from the ECD of the second class (*No single*) for feature `c_total`. This means that the classes *Single* and *No single* are well separable with respect to feature `c_total`. On the other side, the right plot shows that the ECD of the two classes are almost the same for feature `r_noon/day`. As a consequence, we say that `single` is well-separable because there is at least one feature that properly separates the classes.

In terms of class labels, there are natural definitions of class labels for some of the characteristics (e.g., *Single/No single*, or *Family/No family*). For other

9

characteristics (e.g., `age_person`, `#bedrooms`, `floor_area`), we define the class labels (1) according to qualitative considerations gathered during the aforementioned interviews and (2) by adjusting the number and definition of class labels such that each class contains a similar number of households.
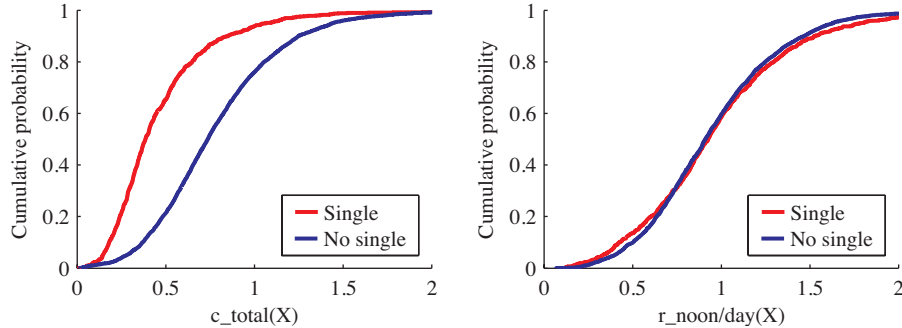

Figure 3: Empirical cumulative distributions of the (unscaled) features `c_total` (left) and `r_noon/day` (right) for characteristic `single`.

4.3. Classifiers

There exist several classifiers that can be used to perform supervised machine learning tasks [32, 33, 34, 35]. These classifiers typically differ in terms of implementation and computational complexity, or in the assumptions they make on the distribution of the data. For the study described in this paper, we have selected five well-known classifiers: the k-Nearest Neighbors (kNN) classifier [32], the Linear Discriminant Analysis (LDA) classifier [32], the Mahalanobis distance classifier [33], the Support Vector Machine (SVM) classifier [34], and the AdaBoost classifier [35].

The right column in table 2 shows that some of the characteristics are imbalanced in the CER data set. This means that some classes have a significantly higher number of samples than other classes. For example, there are 859 households for which the characteristic `single` takes value *Single* and 3,373 for which it takes value *No single*. As we have already outlined in our previous work, this bias affects the performance of some of the classifiers. Since the trained model of these classifiers is biased towards the class with the majority of samples, they often assign samples of the underrepresented classes to the majority class [11]. An effective method to deal with class imbalance consists in *undersampling* the data during the training process [36, 37]. By randomly removing samples from the overrepresented classes, undersampling creates evenly distributed classes (i.e., classes having the same number of samples equal to the number of samples in the smallest class). In order to support applications that rely on identifying samples of underrepresented classes, our system can thus also perform undersampling.

Table 2: List of household characteristics, their class labels, and the number of samples per class in the CER data set. The characteristics eligible for regression are marked with $^{(*)}$.

Characteristic	Description	Classes	No. of samples
age_person$^{(*)}$	Age of chief income earner	Young (age_person < 35) Medium ($35 <$ age_person ≤ 65) High ($65 <$ age_person)	436 2,819 953
all_employed	All adults work for pay	Yes No	1,013 2,409
#appliances$^{(*)}$	Number of appliances	Low (#appliances ≤ 8) Medium ($8 <$ #appliances ≤ 11) High ($11 <$ #appliances)	1,421 1,479 1,332
#bedrooms$^{(*)}$	Number of bedrooms	Very low (#bedrooms ≤ 2) Low (#bedrooms $= 3$) High (#bedrooms $= 4$) Very high ($4 <$ #bedrooms)	404 1,884 1,470 465
cooking	Type of cooking facility	Electrical Not electrical	2,960 1,272
employment	Employment of chief income earner	Employed Not employed	2,536 1,696
family	Family	Family (#adults > 1 and #children > 0) No family	1,118 3,114
floor_area$^{(*)}$	Floor area	Small (floor_area ≤ 100 m^2) Medium (100 $m^2 <$ floor_area and floor_area ≤ 200 m^2) Big (200 $m^2 <$ floor_area)	232 1,198 351
house_type	Type of house	Free (detached or bungalow) Connected (semi-detached or terraced)	2,189 1,964
income$^{(*)}$	Yearly household income	Low (income $< 50,000$) High ($50,000 \leq$ income)	940 997
lightbulbs	Proportion of energy efficient light bulbs	Up to a half About three quarters or more	2,041 2,191
children	Children	Yes (#children ≥ 1) No (#children $= 0$)	1,229 3,003
age_house	Age of building	Old ($30 <$ age_house) New (age_house ≤ 30)	2,151 2,077
#residents$^{(*)}$	Number of residents	Few (#residents ≤ 2) Many ($3 \leq$ #residents)	2,199 2,033
retirement	Retirement status of chief income earner	Retired Not retired	1,285 2,947
single	Single	Single (#adults $= 1$ and #children $= 0$) No single	859 3,373
social_class	Social class of chief income earner according to NRS social grades	A or B C1 or C2 D or E	642 1,840 1,593
unoccupied	Is the house unoccupied for more than 6 hours per day?	Yes No	885 3,347

4.4. Multiple linear regression

Some of the characteristics in table 2 – namely `age_person`, `#bedrooms`, `#appliances`, `floor_area`, `income`, and `#residents` – take values in a continuous interval. For these characteristics we train a regression model in order to estimate the value of the characteristic. We model each of the characteristics individually and use a multiple linear regression model for its simplicity and interpretability of the parameters. The model is expressed as follows:

$$f_R : y_j = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x_j} + \epsilon, \epsilon \sim N(0, \sigma^2), \tag{1}$$

where y_j represents the j-th household's observed value in the training set and $\boldsymbol{x_j}$ denotes the feature vector computed for household j. The coefficients β are then estimated using Ordinary Least Squares regression (OLS) [38].

5. Evaluation process

This section describes how we use the features and classifiers described above to derive quantitative results on the potential to reveal household characteristics from electricity consumption data.

5.1. Performance measures

The first step in determining the performance of a classification outcome is to count the number of correct classifications and the number of misclassifications for each class and thus derive the so-called *confusion matrix* CM. Consider a classification with K classes $(1, ..., K)$ and S samples. The confusion matrix consists of K rows and K columns. The element (i, j) of the confusion matrix represents the number of samples of class i that have been classified as class j. Therefore, the elements on the main diagonal of the matrix CM, indicated as CM_{ii} $(i = 1, ..., K)$, represent the number of correctly classified samples for each class. If $K = 2$, the entries CM_{11}, CM_{22}, CM_{21}, CM_{12} denote the number of *true positives* (TP), *true negatives* (TN), *false positives* (FP), and *false negatives* (FN), respectively.

Sokolova and Lapalme provide an extensive overview of different performance measures for classification tasks [39]. A commonly used performance measure is the *accuracy* of a classifier, which is defined as the total number of the correctly classified samples divided by the total number of samples:

$$ACC = \frac{\sum_{k=1}^{K} CM_{kk}}{S}. \tag{2}$$

We compare the accuracy achieved by the five classifiers considered in this study with the accuracy of two random classifiers. The first is a random guess classifier (RG), which randomly selects a class assuming equiprobable classes. This classifier achieves an accuracy of

$$ACC_{RG} = \frac{1}{K}. \tag{3}$$

To account for the fact that classes are not always equiprobable, we also consider a biased random guess classifier (BRG). The BRG classifier uses knowledge of the proportion of samples of each class in the training data to perform a biased random decision. The accuracy obtained by the BRG classifier is

$$ACC_{BRG} = \sum_{k=1}^{K} (\frac{S_k}{S})^2,$$

(4)

where S_k denotes the number of samples of class k.

The accuracy measure treats all classes equally and is often a weak measure when dealing with imbalanced classes [39, 40]. For this reason, we also utilize the *Matthews Correlation Coefficient* (MCC) to quantify the performance of the considered classifiers [40]. The MCC ranges between -1 and 1, whereas 1 represents a perfect classification, 0 denotes a classification that is no better than a random classification, and -1 shows a disagreement between classification and observation. In case $K = 2$, the MCC is computed as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

(5)

For $K > 2$, we use the generalization of the MCC to multi-class classifications as presented by Gorodkin in [41].

While accuracy and MCC allow one to describe the overall performance of a classifier, utilities are often interested in selecting a specific group of customers, which we call the *target group*, such as the group of households belonging to the class *Single*. To this end, we compute the *true positive rate* (TPR) and *false positive rate* (FPR), which are defined as $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{TN+FP}$ [41]. In the example above, the TPR (or *recall*) indicates the number of correctly estimated *Single* households out of all households that belong to the class *Single*. The FPR indicates how many samples were incorrectly classified as *Single*. The *receiver operating characteristic* (ROC) curve relates these two metrics to each other, illustrating the trade-off between the benefits (true positives) and costs (false positives) of a classification. We implement the method described by Fawcett [42] to create the ROC curve for each target group, or target class, C. The method requires as input the *posterior probability* $P(C|\boldsymbol{x})$ for each sample, which is the probability that a sample belongs to the class C given the feature vector $\boldsymbol{x}$. For $K > 2$, we combine all households that do not belong to the target group into a single group.

To evaluate the performance of the multiple linear regression, we first obtain the estimate $\hat{y}_j$ for each household j as

$$\hat{y}_j = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x_j},$$

(6)

using the parameters $\boldsymbol{\beta}$ and the feature vector $\boldsymbol{x_j}$. We then compare the estimation with the ground truth data y_j by computing the *coefficient of determination* (R^2) as a performance measure [38]:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}.$$

(7)

13

R^2 ranges between 0 and 1 and denotes the proportion of the variance of the estimation error

$$SS_{res} = \sum_j (y_j - \hat{y}_j)^2 \tag{8}$$

to the variance of the ground truth data

$$SS_{tot} = \sum_j (y_j - \bar{y}_j)^2. \tag{9}$$

We further compute the out-of-sample root-mean-square error (RMSE) to evaluate the deviation of the estimation $\hat{y}_j$ to the ground truth data y_j for each of the household characteristics [38].

5.2. Training, evaluation, and feature selection

Listing 1 illustrates the training and evaluation procedure we apply to reveal household characteristics from electricity consumption data. As input, we use a single week of consumption data for all households, which we divide into four disjoint subsets. One subset is used for training the classifiers, the others to validate their performance using a 4-fold cross validation. The performance metrics of interest (accuracy, MCC and ROC curves) are thus computed for each week of the data. In each fold of the cross-validation, a feature selection algorithm determines a subset of the features defined in section 4.1, which is then used by the classifiers. The output of the classifiers is used along with ground truth data to compute the confusion matrix for each classifier. The matrix is then in turn used to compute the performance measures described above. The performance measures are the only difference in this process when performing regression instead of classification.

As line 12 in listing 1 indicates, we rely on the feature selection method SFFS (*sequential floating forward selection* [43]) to determine a suitable set of features $\bar{F} \subseteq F, \bar{F} = \cup_{i=1}^{|F|} c_i f_i, c_i = \{0, 1\}$, where f_i is the i-th feature in F, $|F|$ denotes the size of F, and $c_i = 1$ indicates membership of f_i in $\bar{F}$. There is an optimal set of features F_{opt}, with which a classifier achieves the best value for a specific performance measure (e.g., the highest accuracy). Since F_{opt} typically differs from F [44], feature selection methods approximate F_{opt} by iteratively running the classification (or regression) using different subsets of features. After each run these methods compute a *figure of merit*, which can be any of the performance measures described in the previous section. There are different strategies to maximize the figure of merit and thus optimize the feature set. SFFS is a method that starts with an empty set and consecutively adds the feature to the set that allows one to achieve the highest improvement of the figure of merit. In each step, SFFS also considers removing one or more features from the set, since removing a feature that has been added previously and adding a different one might lead to an increase of the figure of merit. We perform feature selection on the training set as described above. Since the feature selection itself requires both training and test data, we perform another cross-validation on the three subsets of the training set $D \setminus D_i$ in listing 1.

14

Our implementation of SFFS relies on the code provided by the authors of [45]. As an improvement to this existing implementation, we also make the SFFS maintain a logbook of the states it reaches – where a state is represented by a (sub)set of features – in order to prevent it entering infinite loops. To limit the number of iterations and to avoid overfitting, we restrict the removal of features as follows: Assume feature f is added to state s and state $s' = s \cap f$ is reached. A feature $f' \neq f$ is them removed from s' only if the figure of merit of $s'' = s' \setminus f'$ is more than a threshold T higher than the figure of merit of s. We set $T = 0.005$ because, based on our experiments, differences of less than 0.005 in the figure of merit of two states are often due to random effects and thus do not necessarily imply a significant improvement of s'' over s. In this way we avoid overfitting, which could lead to reduced performance when finally using $\bar{F}$ on a new set of data (which we do in lines 13 and 14 of listing 1). Similarly, we limit the number of features selected by SFFS to $|\bar{F}| <= 8$, because adding more than 8 features to $\bar{F}$ often leads to overfitting the feature set to the training data in our experiments.

```
1   In : Consumption data D
2          Classifier Cl
3          Characteristic P, classes C = {C_1, ..., C_n}, n: #classes in P
4   Out: Accuracy ACC_{P,Cl}
5          Matthew Correlation Coefficient MCC_{P,Cl}
6          Posterior probabilities P(C_j|D)_{P,Cl} ∀ C_j ∈ C
7   begin
8          Divide D into disjoint subsets: D = {D_1, D_2, D_3, D_4}
9          for each fold i = 1, ..., 4 do
10             test_set = D_i
11             training_set = D \ D_i
12             F̄ = sffs(training_set)
13             model = train(training_set, F̄, Cl)
14             P(C|D_i) = test(model, test_set, F̄, Cl)
15             CM_i = compute confusion matrix from P(C|D_i) and D_i
16         end for
17         P(C|D)_{P,Cl} = {P(C|D_1), ..., P(C|D_4)}
18         ACC_{P,Cl} = ACC(CM_1, CM_2, CM_3, CM_4)
19         MCC_{P,Cl} = MCC(CM_1, CM_2, CM_3, CM_4)
20  end
```

Listing 1: Evaluation process that is performed for each classifier and characteristic.

5.3. Implementation details

We implemented our system in MATLAB. All results presented in the following section are obtained by performing independent experiments for each of the characteristics listed in table 2, each of the five classifiers (kNN, LDA, Mahalanobis, SVM, and AdaBoost) described in section 4.3, and two different performance measures (accuracy and MCC). For regression, we employ multiple linear regression and use the adjusted R^2 score as a figure of merit. We use the kNN, LDA, Mahalanobis, and AdaBoost classifiers from MATLAB's

Statistics toolbox[2]. For kNN, we choose $k = 5$ and use the Euclidian distance as the distance metric. In case of AdaBoost, we use the AdaBoostM1 and AdaBoostM2 learners to classify characteristics with two and more than two classes, respectively. As for the SVM classifier, we rely on the publicly available implementation $LIBSVM$[3] with a radial basis function kernel.

6. Results

To quantify the performance of our system, we first consider a single week of data in the CER data set separately (week 26). We then repeat our analysis for each week of data, showing that these results are consistent irrespective of which week of data is used. Week 26 was chosen as an "exemplary" week since (1) there is no holiday during this week, (2) the week is not during vacation, and (3) it includes data from households that joined the trial in early 2010. As we show later in this section, experiments on all other weeks of the trial have shown no significant impact (2% standard deviation per characteristic on average) on the performance of our approach. However, by combining the classification results over multiple weeks, the accuracy and the MCC can be improved on average by 3 and 6 percentage points, respectively, compared to a single week analysis.

6.1. Accuracy

Figure 4 shows the accuracy achieved by our system and the two baseline classifiers BG and BRG when estimating the 18 household characteristics of interest. For each characteristic, ACC_{C^*} denotes the highest accuracy value among each of the five classifiers kNN, LDA, Mahalanobis, SVM, and AdaBoost. Showing the accuracy of the best performing classifier allows us to outline the accuracy that can be obtained in principle when using our system for estimating household properties. We leave it to our future work to explore solutions that allow us to maximize the performance of our system in practical scenarios. The results in figure 4 are obtained using the accuracy as a figure of merit during feature selection and without undersampling (i.e., the classifiers have knowledge of the class distribution from the training data). The graph compares the accuracy of our system (ACC_{C^*}, left bars) with the accuracy of the biased random guess classifier (BRG, center bars) and the random guess classifier (RG, right bars). The characteristics on the x-axis are those listed in table 2 in section 4.2. Among the 18 characteristics, there are four three-class characteristics (age_house, #appliances, social_class, floor_area) and one four-class characteristic (#bedrooms). The remaining characteristics are two-class characteristics. For these characteristics, the accuracy of the RG classifier is 33%, 25%, and 50%, respectively. The accuracy of the BRG classifier is computed using equation 4 and the number of samples per class as listed in table 2. If the classes are balanced, ACC_{BRG} is equal to ACC_{RG}. Figure 4 shows that ACC_{C^*}

---

[2]www.mathworks.de/products/statistics
[3]www.csie.ntu.edu.tw/~cjlin/libsvm

exceeds both ACC_{RG} and ACC_{BRG} for all of the 18 characteristics. ACC_{C^*} is 4.0 to 33.8 percentage points higher than ACC_{RG} and 4.0 to 22.3 percentage points higher than ACC_{BRG}. Our system thus outperforms RG and BRG by 22.0 percentage points and 14.2 percentage points, respectively. In terms of individual characteristics, our approach achieves more than 80% accuracy for characteristics single, all_employed, and unoccupied. The worst accuracy of ACC_{C^*} compared to ACC_{BRG} is achieved when estimating the proportion of energy-efficient light bulbs (lightbulbs) in a household. In this case, ACC_{C^*} exceeds ACC_{BRG} by only 4.0 percentage points. We believe this results from the fact that the mere number of lightbulbs is not reflected in the electricity consumption. We expect this result to improve when classifying the actual usage of energy-efficient light bulbs. However, this data is not available in the CER data set.
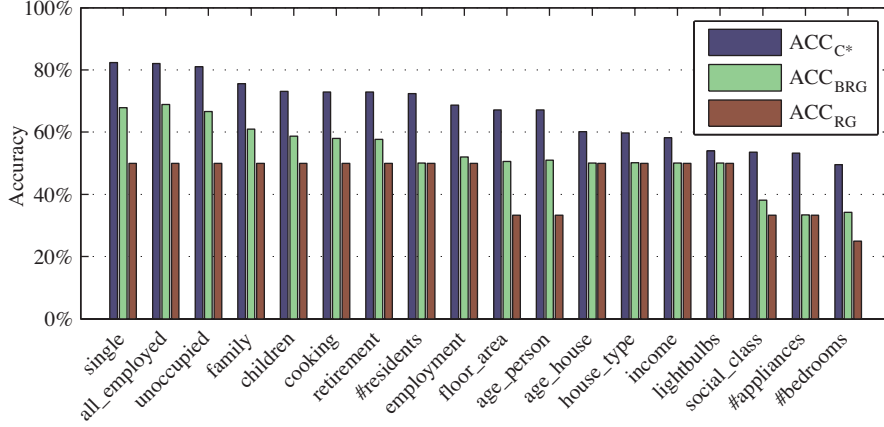

Figure 4: Accuracy of our system (ACC_{C^*}) compared to the random guess (ACC_{RG}) and biased random guess (ACC_{BRG}) classifiers based on week 26 of the trial.

Figure 5 illustrates the accuracy of all five classifiers. For most of the characteristics, the difference between the highest and lowest accuracy is 10 percentage points or less. The SVM classifier achieves the highest accuracy for 13 of the 18 characteristics. The AdaBoost classifier performs 1 percentage point worse than the SVM classifier on average. The LDA and Mahalanobis classifiers show similar performance for some of the characteristics but have a low accuracy for the characteristics with imbalanced classes (e.g., floor_area, #bedrooms). Among the considered classifiers, SVM is thus the one providing the overall best performance in terms of accuracy. If our system were to be used to maximize the estimation accuracy, we would thus use SVM as the default classifier. We leave it to our future work to verify whether this consideration can be generalized to other data sets.
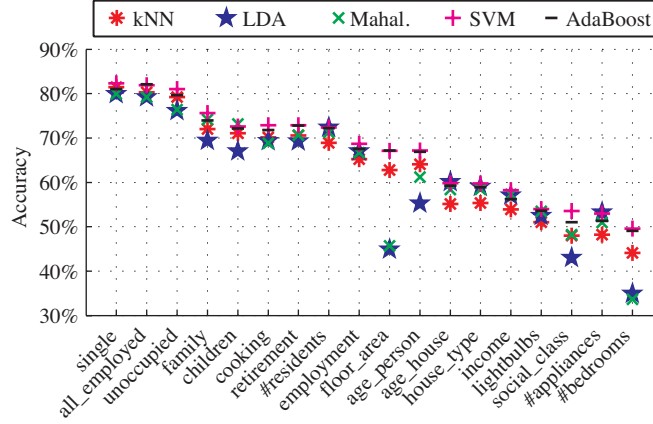
Figure 5: Accuracy of different classifiers for all household characteristics. The classifiers are trained using data from week 26 of the CER data set.

6.2. Matthews Correlation Coefficient

As described in section 5.1, the Matthews Correlation Coefficient (MCC) is typically a more suitable performance measure than accuracy when classes are imbalanced. The MCC "rewards" true positives of the underrepresented classes and thus "punishes" classifiers that bias their model too strongly towards the overrepresented classes. It is a performance measure that ranges between -1 (i.e., total disagreement between the ground truth data and the estimation) and 1 (total agreement), whereas the MCC is 0 for random estimations [46]. Figure 6 shows the MCC for each of the classifiers considered in our system. In these experiments, we used the MCC as a figure of merit during feature selection and use undersampling to prevent classifiers from biasing their model towards the overrepresented classes. The plot shows that for characteristics related to the number of people in a household (i.e., `single`, `#residents`, `family`, `children`), the considered classifiers achieve high values of MCCC (up to 0.459). Figure 6 further shows that the MCC is also high (up to 0.346) for characteristics related to occupancy (i.e., `employment`, `all_employed`, `unoccupied`, `retirement`). Finally, classifying characteristic `#appliances` provides an MCC of 0.31. The results indicate that for these characteristics, classification is feasible. Whether the results are good enough to provide energy efficiency services must be decided on a per-application basis and is left to future work.

With respect to the individual classifier performance, the kNN classifier performs worse than the LDA, Mahalanobis, SVM, and AdaBoost classifiers across (almost) all of the characteristics. Among the other four classifiers, neither of the four classifiers' performances dominates over all characteristics, which makes the choice of classifier dependent on the characteristic to be classified. We thus argue for the use of a comprehensive system that uses a particular classifier depending on the specific characteristic to be estimated.
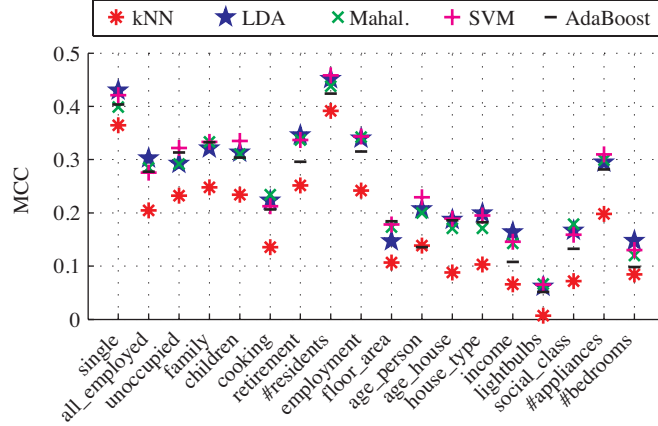
18

Figure 6: Matthews Correlation Coefficient (MCC) of different classifiers for all household characteristics. The classifiers are trained using data from week 26 of the CER data set.

6.3. ROC curves

We compute the ROC curves on the basis of the experiments that include undersampling and use the MCC as a figure of merit during feature selection. We chose this particular configuration over the one that optimizes for accuracy, because based on our experiments the latter one results in poor performance when targeting the underrepresented class of a characteristic. Figure 7 shows ROC curves for different characteristics and target classes. The title of each subplot indicates the characteristic and the target class as defined in table 2. In case of multi-class characteristics, we combine the non-target classes to achieve a two-class setting. The diagonal, dashed line denotes the ROC curve of the RG classifier. The other lines show the ROC curves of the kNN, LDA, Mahalanobis, SVM, and AdaBoost classifiers. For each class and classifier, the graphs show the true positive rate (TPR, y-axis) that can be achieved at a given false positive rate (FPR, x-axis) and vice versa. Each point on the ROC curve represents a household h with a posterior probability $P = P(C_i|\boldsymbol{x_h})$, which indicates the confidence that h belongs to class C_i given its feature vector $\boldsymbol{x_h}$. All points on the ROC curve on the right side of this point correspond to households with a higher posterior probability compared to P.

In practice, the ROC curves enable utilities to choose an FPR (or TPR) λ when selecting customers of a certain class. Depending on λ, the point $p = ROC(\lambda)$ on the curve implicitly defines a subset of households that exhibits an FPR (or TPR) of λ. These are all households that have a higher posterior probability than the household at p. The number of households in the subset depends on the selection of λ. Figure 7 shows that when selecting single-person households (subplot with title "single: yes"), for instance, a utility can identify 50% of all single-person households with only 10% false positives in the selected set using the LDA classifier (blue line with '*' markers). Thus, if 50% of false positives is acceptable for the application envisioned by the utility (e.g.,

a marketing campaign), it can select a threshold such that the resulting subset of households contains 90% of the single-person households. Overall, for 11 of the 18 groups, it is possible to identify 50% of the households of a specific group with a false positive rate lower than 20%. However, the plot also illustrates that selecting households with high incomes or households that have a large share of non-energy-saving lightbulbs are difficult to identify, as their ROC curve is close to the diagonal curve in the plot.

6.4. Regression

The regression model described in section 4.4 allows us to estimate continuous values of selected characteristics instead of assigning the household to pre-defined, discrete classes. Figure 8 shows the results for characteristics `age_person`, `#bedrooms`, `#appliances`, `floor_area`, `income`, `#residents`. The x-axis of each plot shows the ground truth values for the characteristics. The reason for choosing a box plot instead of a scatter plot for some of the characteristics is that the ground truth data is binned (e.g., `age_person`, `income`) or provides only a few discrete values (e.g., `#bedrooms`, `#residents`). The y-axis shows the estimation obtained by applying f_R from equation 1 to the features of the test data. The subplot at the bottom right, for instance, shows one box per group of households with 1, 2, 3, 4, or 5+ (5 or more) residents. The red lines denote the median of the estimated number of residents for each of the five groups, and the top and bottom ends of the boxes denote the 25th and 75th percentiles, respectively. For the characteristic `floor_area`, each value is plotted individually in a scatter plot.

For each of the characteristics, the figure shows the root-mean-square error (RMSE) and the coefficient of determination (R^2). The latter one ranges between -1 and 1, whereas 0 shows no correlation between the estimated and the ground truth data and 1 indicates a perfect estimation. The characteristics `#residents` and `#appliances` achieve the highest R^2 with 0.30 and 0.29, respectively. The characteristics `#bedrooms`, `age_person`, and `floor_area` follow suit with 0.14, 0.17, and 0.14, respectively. Finally, `income` is hardest to reveal with a very low R^2 of 0.083. Although the plots show a clear correlation between the estimated and the actual values, the R^2 score is overall relatively low. We assume this is due to the fact that the linear regression model is very sensitive to outliers. Examples of such outliers are households that have their ground truth incorrectly specified in the questionnaires. These results suggest that utilities should rely on the estimated class rather than striving for exact, continuous values. As a part of our future work, we aim at improving the R^2 scores and thus the applicability of the regression analysis for utilities by automatically identifying households or groups of households that negatively affect the performance of the analysis.

6.5. Stability of the results

The results presented thus far are based on the analysis of a single week of consumption data (week 26). We then extended the experiments to the whole
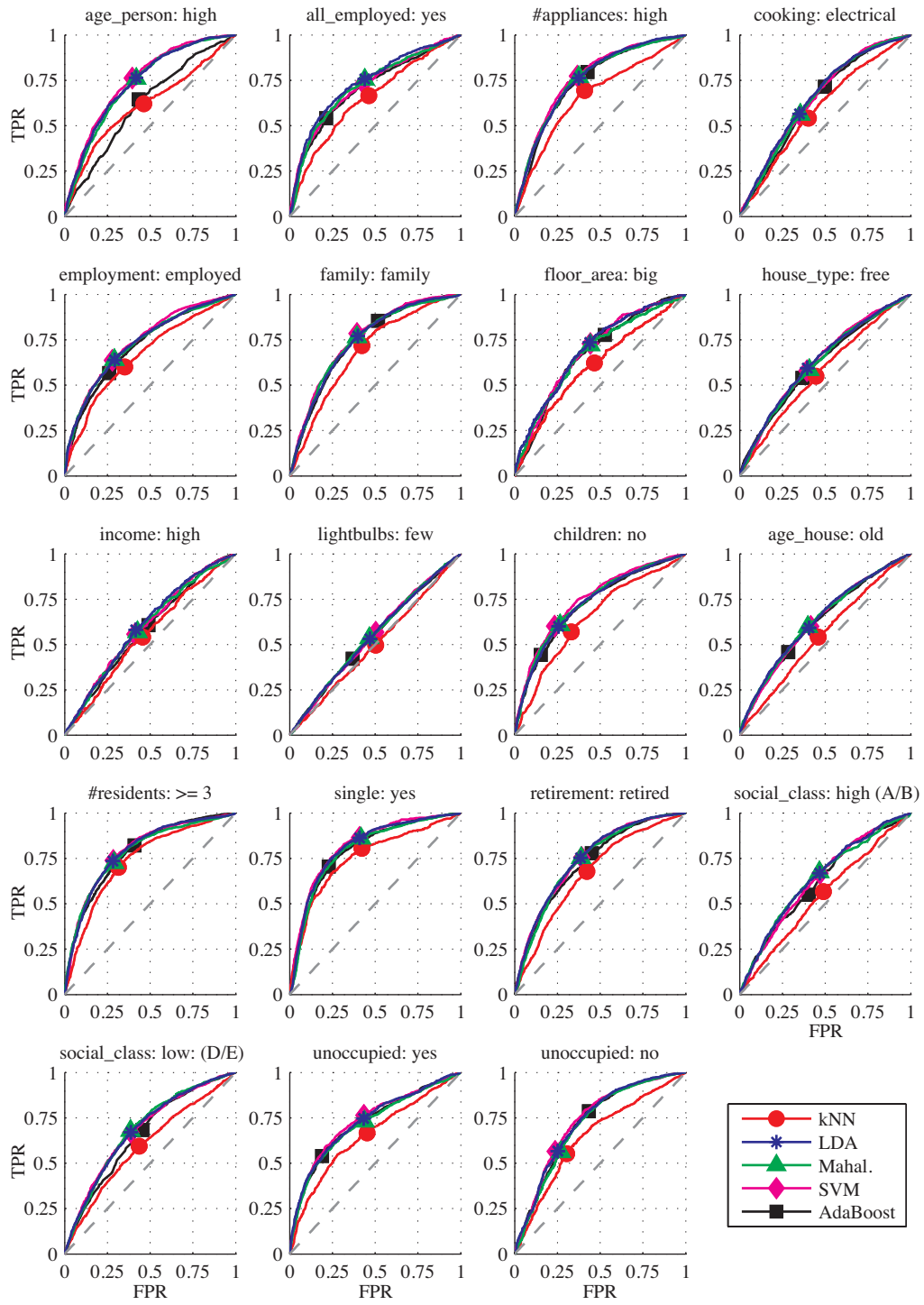
Figure 7: ROC curves that show the trade-off between true positive rate and false positive rate for multiple characteristics. Each subplot compares the five classifiers (colored lines) with the random guess (dashed line). The title of each subplot describes the characteristic and the target class.
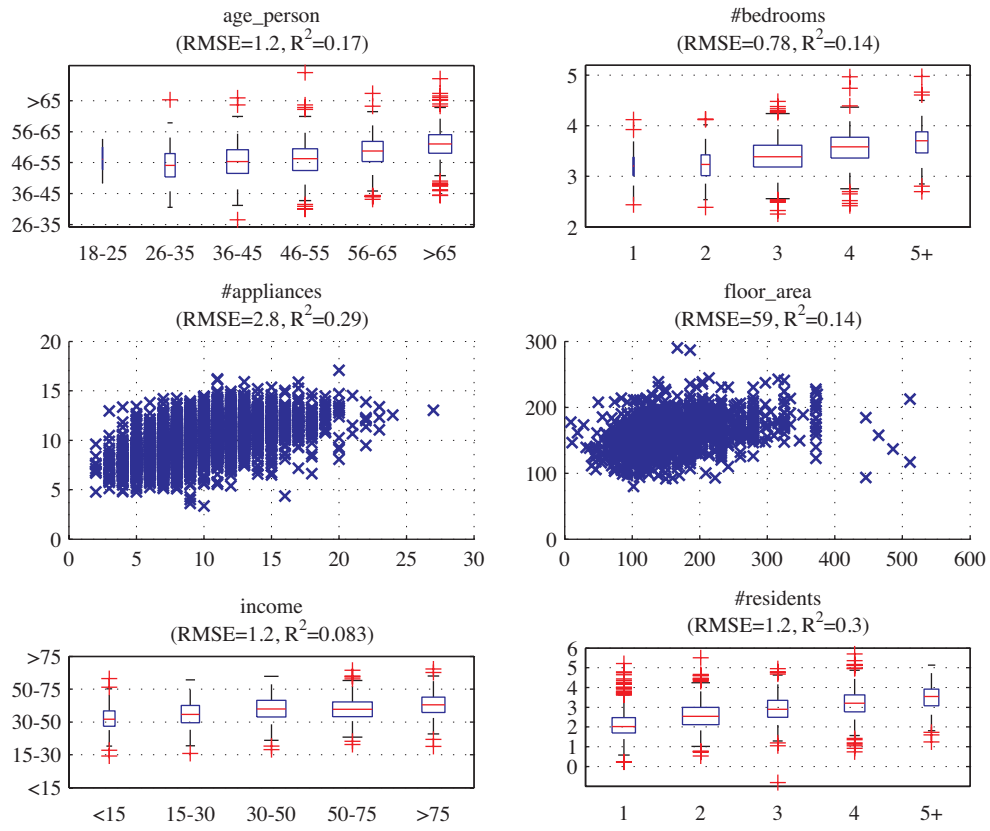
Figure 8: Regression analysis for selected characteristics and consumption data of week 26.

data set by classifying all of the characteristics on each week of the study, following the procedure described in listing 1. This means we evaluate each week separately by computing the features on this particular week and (as before) using different households for training and testing. For these experiments, we relied on the LDA classifier only, because it is the fastest of the five classifiers evaluated in this paper. We assume the results can be generalized to the other classifiers, which is subject to future testing. Figure 9 shows the results of these experiments. The left plot illustrates the accuracy for each characteristic on each week encoded with colors ranging from dark blue (30% accuracy) to dark red (80% accuracy). Similarly, the right plot shows the MCC ranging from dark blue (0) to dark red (0.5). The plots show that the difference between weeks is relatively low for most of the characteristics: In terms of accuracy, the average standard deviation for all characteristics is 0.017. Only the characteristics `floor_area` and `#appliances` exhibit large variations with a standard deviation of 0.034 and 0.033, respectively. For these two characteristics, for instance, the difference between minimum and maximum accuracy is 13.7% and 18%, respectively, whereas for all other characteristics this difference is 6% on average. For the MCC, the standard deviation is below 0.04 for all characteristics. An interesting observation can be made for weeks 50 to 57 of the trial, where the classification performs slightly worse than for the rest of the weeks. This is particularly true for characteristics that are related to the number of persons in the household (e.g., `single`, `#residents`) or related to the presence of people (e.g., `retirement`, `employment`). We believe that this loss in performance is due to the fact that these weeks represent summer vacation in Ireland. Thus, in these weeks we observe non-usual consumption patterns, which cause the classification to perform less reliably than in other weeks.

The experimental results reported above show that it is possible to reliably extract household characteristics from electricity consumption data irrespective of the specific week of data used for the analysis. This implies that utilities might need to collect as little as one week of fine-grained data (i.e., one sample every 30 minutes) to be able to extract household characteristics. In future work, we plan to investigate whether the results are also promising when we train the classifiers using a specific week of data and classify (different) households using a different week of data. This would allow utilities to include new households in the analysis (e.g., households with a newly installed smart meter) without retraining the classifiers. Next, the stability of the results shows that utilities can build weekly customer profiles using our approach. For each customer, such a profile can for instance show when a family grows or goes on vacation. The fact that results are stable across multiple weeks further enables us to identify "atypical" weeks for individual households as outliers (e.g., when the house is unoccupied because the family is on vacation). In the following section, we show that running our analysis on multiple weeks allows us to compensate for such outliers and improve the performance of the classification.
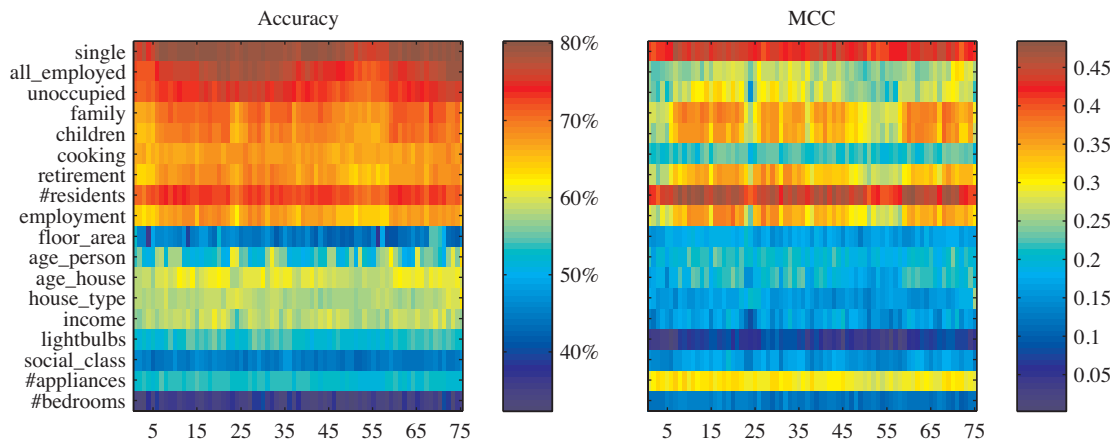
Figure 9: Classification results for each of the characteristics on all 75 weeks of the trial using the LDA classifier – accuracy (left), MCC (right).

6.6. Multiple weeks

In this section, we combine the results of the classifiers over multiple weeks. We train the classifiers for each week separately using the households in the training set. Next, we estimate – for each household in the test set – the class of the household for each week and then assign the household to the class C that was estimated most often. We decided for this majority vote for its simplicity; we leave alternative methods of performing the analysis on multiple weeks (e.g., computing the features over a longer time period) for future work. Ultimately, we create a confusion matrix for each of the characteristics by comparing C with the actual ground truth. As the number of households varies over the weeks due to missing meter readings for some households, we only consider households in the analysis for which data over more than 50 weeks is available. Table 3 lists both accuracy and MCC of the classification using week 26 only (as described in detail above) and the results obtained when running the analysis over the whole period of the trial (75 weeks). The table shows that the accuracy increases by up to 10 percentage points (3 percentage points on average) and the MCC by up to 10 percentage points (6 percentage points on average). The performance of the classification only decreases for characteristic house_type; however this change is very low (1 percentage point for the MCC). We thus argue that, in practical scenarios, utilities interested in maximizing classification performance should utilize several weeks of data to estimate household characteristics. However, the computational effort increases linearly with the number of weeks. It is a part of our future work to investigate this trade-off in more detail.

6.7. Discussion and limitations

The experimental results presented thus far show that three types of characteristics can be inferred particularly well from electricity consumption data.

Table 3: Accuracy and MCC for each characteristic obtained by assigning to each household in the test set the majority of classifications over the whole trial ("vote"). The results are based on the LDA classifier.

	Accuracy		MCC	
Characteristic	Week 26	Vote	Week 26	Vote
single	80%	82%	0.43	0.5
all_employed	79%	79%	0.3	0.32
unoccupied	76%	76%	0.29	0.38
family	69%	74%	0.32	0.42
children	67%	73%	0.31	0.41
cooking	69%	71%	0.22	0.29
retirement	69%	73%	0.35	0.43
#residents	72%	76%	0.45	0.51
employment	67%	72%	0.34	0.44
floor_area	45%	50%	0.15	0.21
age_person	55%	59%	0.21	0.3
age_house	60%	64%	0.19	0.28
house_type	59%	59%	0.2	0.19
income	57%	61%	0.16	0.23
lightbulbs	52%	55%	0.062	0.098
social_class	43%	53%	0.17	0.22
#appliances	53%	56%	0.29	0.34
#bedrooms	35%	39%	0.15	0.15
Mean	**62%**	**65%**	**0.26**	**0.32**

These are characteristics that reflect the occupancy state of the house (e.g., employment, unoccupied), the number of persons in the house (e.g., single, #residents, family), and the appliance stock (#appliances). On the other hand, characteristics related to the dwelling itself (e.g., floor_area, #bedrooms) are more difficult to extract from electricity consumption data. This is due to the fact that heating and cooling, which is typically reflected in the consumption data (see section 2), is not included in the consumption data available for this study. The results show that the income of a household is also difficult to infer from electricity consumption data.

It is in general important to note that the results presented in this paper might be affected by inaccuracies in the ground truth data. Questionnaire answers given by the participants in the CER study can be wrong, ambiguous, or based on estimations. For characteristic all_employed, for example, the questionnaires do not specify full-time or half-time employment. Characteristic unoccupied relies on the estimated absence rather than on actual measurements. For characteristic income, the process of extracting well-separated classes from the ground truth data was difficult due to the complex structure of questions that captures the income of the respondents. For instance, they could specify their income on a yearly or monthly basis as well as before or after tax according to the questionnaire.

A major challenge to applying this work in practice is to collect reliable ground truth data. This step typically requires surveys, which are costly and cumbersome to perform. Yet, even if only a small percentage of customers

reply [6], this small amount can be used to train the classifiers and estimate the characteristics of the remaining households. There is also the possibility to use ground truth data from a different data set (e.g., collected from a project performed in a different geographic region), which we plan to evaluate in future work.

At the moment, we evaluate each of the five classifiers and then decide which one performs best. To implement the approach in real scenarios, however, we must decide on the classifier on the basis of the training data only. Similarly to what we do for feature selection, we propose dividing the training data into two sets and using one of the sets to train the classifiers and the other set to evaluate their performance. This process also allows for fine-tuning of each of the classifiers for each of the characteristics. As this paper merely provides an overview of the potential of large-scale electricity consumption data analysis, we leave it for future work to optimize the classifiers accordingly.

7. Implications for households, utilities, and policy makers

The approach presented in this paper automatically reveals household characteristics including the number of inhabitants, the size of the property, or the number of appliances by analyzing households' electricity consumption. It utilizes data from smart electricity meters that capture consumption information at 30-minute intervals and thus can be used in combination with almost all smart meters that are currently being rolled out throughout the world. On the basis of these findings, we see several (positive and negative) implications for households, utilities, and policy makers, which are outlined below.

7.1. Households

Households can benefit from more informative and precise, more enjoyable, and motivating energy efficiency campaigns: As the analysis provided in this paper allows for identifying energy-relevant household characteristics, it becomes possible for utilities to benchmark households based on similar demographics and household type [4, 47]. It also enables assessing a household's energy efficiency using an energy-efficiency label that is easy to understand. Utilities can further group similar households for engagement campaigns, summarize information or provide tips and tricks that are relevant for each group, or define peer households to realize concepts based on games to increase user engagement. It has been shown in the past that efficiency campaigns benefit from tailoring advice and motivational cues to the recipient [3, 4, 5]. Overall, targeted campaigns might help to make efficiency-related topics interesting and therefore win the attention of more households. Ultimately, information can be directed in a way that triggers savings in terms of both electricity and money.

At the same time, the findings have strong implications for consumer privacy as well: Our system makes it possible to extract information that consumers may prefer to keep private, including data related to income, employment status, status of the relationship, or social class. Thus, households should engage in

a discussion with those who capture and want to use the data, urging them to make techniques for privacy protection an inherent feature of the emerging smart metering infrastructure [48].

7.2. Utilities

Utilities benefit from insights into customers that they can reveal using the proposed system to estimate household characteristics. The same information that helps to make energy saving campaigns more effective can help to better market products and services. The latter includes identifying target households for specific offerings, e.g., promoting solar panels only to mid to high-income customers who live in a house rather than in an apartment and offering green tariffs preferentially to families with young children. The revealed household characteristics might also help to lower the cost of efficiency campaigns by targeting households with a high potential such as those that show a mismatch between household characteristics and energy demand. Other actions that become possible include tailoring behavioral campaigns to retired individuals or young professionals, or to concentrate on behavioral campaigns rather than on triggering investments in low-income households [2]. Utilities we have collaborated with [10] also hypothesize that more directed customer interaction and better savings advice will improve customer satisfaction and ultimately customer retention. The latter is relevant in particular in energy markets that are liberalized or that face liberalization such that customers can freely choose their utility [49]. In short, the customer insight gained by applying the proposed techniques helps utilities to better allocate their budget, offer directed savings advice, and ultimately boost the impact of their sales and efficiency campaigns.

7.3. Policy makers

Given the advent of the outlined approach, policy makers need to define the rules that govern the use of metering data. It is crucial to promote the beneficial effects including increased energy efficiency and more targeted energy consulting services, and yet to limit the undesired uses of these techniques. As it is at this stage still unclear what "undesired uses" comprises, the stakeholders need to investigate the utilities' and private individuals' interests and find a compromise [50]. Whereas the former probably have an interest in leveraging the retrieved information for marketing campaigns of all sorts, private persons may demand varying levels of privacy protection [48] – depending on the culture they are embedded in.

Consequently, policy makers need to strike a balance between a regime that allows the full materialization of the benefits of smart meter data analytics and regulation that entirely protects privacy. Here, it is important to define who can access the data and what it is used for. One viable approach is to let individual households decide who has access and what they can do with the data [51]. Alternatively, it is technologically also feasible to design solutions in which data is not sent to the utility but rather processed within the household. Households that provide their data for training purposes can then be rewarded in such a

model. However, applications that can be realized by such an approach are still subject to research. Beyond technical solutions, it is also possible to ensure that the information can be used at a large scale by designing an opt-out regime that by default grants utilities access to the data unless an individual actively decides against it. Experience shows that the number of customers who opt out is relatively low, as would be the number of individuals who actively decided for making their data available in an opt-in regime [52].

8. Conclusion and outlook

Alongside the adoption of smart metering systems, data analytic techniques will emerge to make use of the data beyond the realization of dynamic tariffs, consumption visualization, and efficient meter-to-cash processes. The work presented here can help experts to develop tools that allow them to gain important insights into customers, to provide personalized and scalable energy-efficiency programs and at the same time provide managers and policy makers with a precise idea of what can be done with smart metering infrastructure in the near future. From our perspective, the results are both fascinating from a technical perspective and also raise concerns given their privacy implications. By showing what will become possible at a very large scale and how it can be done, we aim to trigger future research in the technical domain as well as intensify the discussion on the rules and regulations that will be needed to govern the data analytics. In particular, we see future technical work especially in the fields of utilizing correlations between characteristics to further improve the performance of the estimation, the increase of the number of household characteristics that can be identified, and approaches to protect consumer privacy. Here it is particularly interesting to evaluate what can be achieved with more finely grained data (e.g., one measurement per minute or even per second) and what the performance loss is with daily, monthly, or even yearly measurements. We are also investigating potential applications for utilities on the basis of the results provided in this paper. Managers at utilities will have to answer the question how to exactly use the newly-won customer insights for improved energy efficiency programs to increase customer satisfaction and customer retention. For policy makers, privacy issues will become more pressing, and the debate needs to be extended given the very detailed information that can be retrieved when smart metering systems are installed.

References

[1] W. Poortinga, L. Steg, C. Vlek, G. Wiersma, Household preferences for energy-saving measures: A conjoint analysis, Journal of Economic Psychology 24 (1) (2003) 49–64.

[2] I. Vassileva, J. Campillo, Increasing energy efficiency in low-income households through targeting awareness and behavioral change, To appear in Renewable Energy.

[3] H. Allcott, Social norms and energy conservation, Journal of Public Economics 95 (9) (2011) 1082–1095.

[4] I. Ayres, S. Raseman, A. Shih, Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage, Tech. Rep. w15386, National Bureau of Economic Research (2009).

[5] N. J. Goldstein, R. B. Cialdini, V. Griskevicius, A room with a viewpoint: Using social norms to motivate environmental conservation in hotels, Journal of Consumer Research 35 (3) (2008) 472–482.

[6] I. A. Stoop, The hunt for the last respondent: Nonresponse in sample surveys, Sociaal en Cultureel Planbu, 2005.

[7] J. Vasconcelos, Survey of regulatory and technological developments concerning smart metering in the European Union electricity market, EUI RSCAS Policy Papers (1) (2008) 1–64.

[8] Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009 concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC.

[9] Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC.

[10] C. Beckel, L. Sadamori, S. Santini, Towards automatic classification of private households using electricity consumption data, in: 4th ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings (BuildSys), ACM, 2012, pp. 169–176.

[11] C. Beckel, L. Sadamori, S. Santini, Automatic socio-economic classification of households using electricity consumption data, in: 4th International Conference on Future Energy Systems (e-Energy), ACM, 2013, pp. 75–86.

[12] P. Bertoldi, S. Rezessy, E. Lees, P. Baudry, A. Jeandel, N. Labanca, Energy supplier obligations and white certificate schemes: Comparative analysis of experiences in the European Union, Energy Policy 38 (3) (2010) 1455–1469.

[13] J. L. Sousa, A. G. Martins, H. Jorge, Dealing with the paradox of energy efficiency promotion by electric utilities, Energy 57 (2013) 251–258.

[14] M. Zeifman, K. Roth, Nonintrusive appliance load monitoring: Review and outlook, in: International Conference on Consumer Electronics (ICCE), IEEE, 2011, pp. 239–240.

[15] A. Zoha, A. Gluhak, M. A. Imran, S. Rajasegarar, Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey, Sensors 12 (12) (2012) 16838–16866.

[16] K. Carrie Armel, A. Gupta, G. Shrimali, A. Albert, Is disaggregation the holy grail of energy efficiency? The case of electricity, Energy Policy 52 (2012) 213–234.

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

[18] J. Kwac, C.-W. Tan, N. Sintov, J. Flora, R. Rajagopal, Utility customer segmentation based on smart meter data: Empirical study, in: 4th International Conference on Smart Grid Communications (SmartGridComm), IEEE, 2013, pp. 720–725.

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

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

[21] A. Albert, R. Rajagopal, Smart meter driven segmentation: What your consumption says about you, IEEE Transactions on Power Systems 28 (4) (2013) 4019–4030.

[22] I. Sánchez, I. Espinós, L. Moreno Sarrion, A. López, I. Burgos, Clients segmentation according to their domestic energy consumption by the use of self-organizing maps, in: 6th International Conference on the European Energy Market (EEM), IEEE, 2009, pp. 1–6.

[23] T. Räsänen, J. Ruuskanen, M. Kolehmainen, Reducing energy consumption by using self-organizing maps to create more personalized electricity use information, Applied Energy 85 (9) (2008) 830–840.

[24] J. Kolter, J. Ferreira, A large-scale study on predicting and contextualizing building energy usage, in: 25th Conference on Artificial Intelligence (AAAI), AAAI Press, 2011.

[25] F. McLoughlin, A. Duffy, M. Conlon, Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study, Energy and Buildings 48 (2012) 240–248.

[26] F. McLoughlin, Characterising domestic electricity demand for customer load profile segmentation, Ph.D. thesis, Dublin Institute of Technology (2013).

[27] 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.

[28] T. Kohonen, Self-organizing maps, 3rd Edition, Springer, 2001.

[29] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems 2 (1) (1987) 37–52.

[30] J. Osborne, Notes on the use of data transformations, Practical Assessment, Research & Evaluation 8 (6) (2002) 1–8.

[31] M. C. Wang, B. J. Bushman, Using the normal quantile plot to explore meta-analytic data sets, Psychological Methods 3 (1) (1998) 46–54.

[32] C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.

[33] T. W. Anderson, An introduction to multivariate statistical analysis, John Wiley & Sons, 1984.

[34] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

[35] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: 2nd European Conference on Computational Learning Theory (EuroCOLT), Springer, 1995, pp. 23–37.

[36] N. Japkowicz, The class imbalance problem: Significance and strategies, in: International Conference on Artificial Intelligence (ICAI), CSREA Press, 2000, pp. 111–117.

[37] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284.

[38] J. O. Rawlings, S. G. Pantula, D. A. Dickey, Applied regression analysis: A research tool, Springer, 1998.

[39] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (4) (2009) 427–437.

[40] B. W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure 405 (2) (1975) 442–451.

[41] J. Gorodkin, Comparing two K-category assignments by a K-category correlation coefficient, Computational Biology and Chemistry 28 (5) (2004) 367–374.

[42] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.

[43] P. Pudil, F. Ferri, J. Novovicova, J. Kittler, Floating search methods for feature selection with nonmonotonic criterion functions, in: Pattern recognition, 1994. Vol. 2 - Conference B: Proceedings of the 12th IAPR International Conference on Computer Vision & Image Processing, Vol. 2, IEEE, 1994, pp. 279–283.

[44] A. Whitney, A direct method of nonparametric measurement selection, IEEE Transactions on Computers 100 (9) (1971) 1100–1103.

[45] S. Theodoridis, A. Pikrakis, K. Koutroumbas, D. Cavouras, Introduction to pattern recognition: A matlab approach, Elsevier, 2010.

[46] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: An overview, Bioinformatics 16 (5) (2000) 412–424.

[47] M. Chetty, D. Tran, R. E. Grinter, Getting to green: Understanding resource consumption in the home, in: 10th International Conference on Ubiquitous Computing (UbiComp), ACM, 2008, pp. 242–251.

[48] E. McKenna, I. Richardson, M. Thomson, Smart meter data: Balancing consumer privacy concerns with legitimate applications, Energy Policy 41 (2012) 807–814.

[49] A. Payne, P. Frow, Relationship marketing: Key issues for the utilities sector, Journal of Marketing Management 13 (5) (1997) 463–477.

[50] S. R. Rajagopalan, L. Sankar, S. Mohajer, H. V. Poor, Smart meter privacy: A utility-privacy framework, in: 2nd International Conference on Smart Grid Communications (SmartGridComm), IEEE, 2011, pp. 190–195.

[51] E. L. Quinn, Smart metering and privacy: Existing laws and competing policies. A report for the Colorado public utilities commission (2009).

[52] E. J. Johnson, D. Goldstein, Do defaults save lives?, Science 302 (5649) (2003) 1338–1339.