# Towards Automatic Classification of Private Households Using Electricity Consumption Data

Christian Beckel
Institute for Pervasive Computing
ETH Zurich
8092 Zurich, Switzerland
beckel@inf.ethz.ch

Leyna Sadamori, Silvia Santini
Wireless Sensor Networks Group
TU Darmstadt
64283 Darmstadt, Germany
{sadamori,santinis}@wsn.tu-darmstadt.de

## Abstract

The ongoing liberalization of the energy market makes energy providers increasingly look at premium services – like personalized energy consulting – as preferred methods to bind existing customers and attract new ones. Providing such services, however, requires knowledge of specific properties of the customer's household – like its size and the number of persons living in it. In this paper, we investigate how such properties can be inferred from the fine-grained electricity consumption data provided by digital electricity meters. In particular, we focus on exploring which properties are both interesting and likely to be identified using well-known classification methods. To this end, we first elicit a set of interesting properties by performing in-depth interviews with employees of three different energy providers. We then explore a large set of electricity consumption traces using a self-organizing map. This analysis allows to identify a set of household properties that are likely to be inferable from electricity consumption data using standard classification methods. For instance, our results show that the size of a household and the income of its occupants are properties that are both highly useful to energy providers as well as likely to be detectable using an automatic classification system.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms

Design, Algorithms, Management

### *Keywords*

Data mining, self-organizing maps, smart electricity meters, energy consumption analysis



*Buildsys'12*, November 6, 2012, Toronto, ON, Canada.


## 1 Introduction

In a 2006 directive the European Union encourages its member states to support the deployment of individual electricity meters "*that accurately reflect the final customer's actual energy consumption and that provide information on actual time of use*" [5]. This measure is part of an encompassing strategy to improve energy efficiency and liberalize the energy market in European countries. Digital electricity meters have indeed been already deployed in millions of households in Italy and Sweden [18] and energy providers in Switzerland, Germany and other also non-European countries are following suit [6]. By providing a detailed report about electricity consumption, digital electricity meters enable providers to bill their customers according to dynamic pricing policies. In future scenarios, electricity prices might change as frequently as every 15 minutes and thus measurements of electricity consumption that are at least as frequent are required in order to correctly bill consumers. This can in turn leave providers with large amounts of fine-grained data about their customers' electricity consumption. In the past few years, several authors investigated potential benefits that could arise – for both consumers and providers – by using this data for other goals than billing. Several studies have for instance investigated load disaggregation methods to provide a "per-device" consumption feedback to the customer [20, 11]. Others aimed at identifying consumption profiles with similar temporal behavior in order to derive better peak-load prediction algorithms [14, 13].

In this paper we describe how specific properties of a household – like its size or the number of individuals living in it – can be inferred from its electricity consumption profile. By computing adequate features of the consumption curve and using a well-known clustering technique we are able to show that it is possible to roughly separate groups of customers according to selected properties. The ability to derive additional knowledge about their customers using only electricity consumption data is of particular interest to providers. Indeed, this knowledge can enable electricity providers to offer new, better, or more customer-tailored energy services. As the ongoing liberalization of the energy market makes it easier for customers to switch their providers, a comprehensive offer of such "premium" services might represent a competitive advantage to bind existing customers and attract new ones. As we elicit through in-depth interviews with employees of three different medium-size Swiss energy providers, an important example of such additional services is represented by energy consulting. Customers of several European energy providers can already request such a consulting service in order to survey their household and identify ways to reduce their energy consumption.This can often be achieved simply by replacing inefficient devices but also by discovering incorrect wirings that can make a customer pay for its neighbor's hyperactive boiler. As a consequence of the current political pressure towards an overall more thrifty usage of energy resources, providers are also starting to offer energy consulting services for free to their customers. In this scenario, the more the provider knows about its customers, the better it can select those most likely to be interested in a consulting session and those that can gain the most out of it. As also elicited through our interviews, however, energy providers often know very little about their private customers. Gathering additional information about customers' households using only the already available consumption data can thus significantly improve the ability of a provider to offer valuable premium services to its customers. Clearly, in order to use the electricity consumption measurements to other goals than billing, energy providers will need to carefully inform their customers and request the possibly necessary permissions. Nonetheless the possibility to automate the process of discovering significant properties of a household remains a very attracting solution with respect to – or even in combination with – customer surveys.

To investigate the possibility to automatically classify private households using traces of electricity consumption data, we performed a preliminary study using traces from more than 3,000 private households. The traces contain measurements at a 30-minute granularity and have been collected for a period of more than 1.5 years. Ground-truth information about the properties of the households – including household size, number of occupants, type of heating, etc. – is also available. The complete data set has been collected in the context of a smart metering study conducted by the Irish Commission for Energy Regulation (CER) and has been recently made available to the public.[1] The availability of such a rich and large data set makes it possible – to the best of our knowledge for the first time – to investigate in a quantitative way the specific use of the electricity consumption data described in this paper.

In order to cast our problem as a standard classification task, however, the definition of appropriate classes, i.e., of appropriate properties of the households that should be inferred from the data, is first required. The definition of these classes can for instance be done depending on the intended use of the classification outcome or on the actually available data. We refer to these two approaches as the *application-driven* and *data-driven* class definition methods, respectively. In the first case, we consider a target application and define a number of properties of the households that are relevant in order to support it. In the context of this paper, we target application is the energy consulting service described above. By analyzing the results of the mentioned interviews we thus elicit a first significant set of household properties – and thus classes of our classification problem. We then turn to a data-driven approach and analyze the CER data set. In particular, we define a number of potentially interesting features of the consumption profiles and investigate their correlation to several different household properties. To this end, we rely on so-called self-organizing maps [9]. Our results show that, for instance, the size of a household and the income of its occupants are both interesting for the energy consulting application and most likely to be reliably identified by a common classifier.

The remainder of this paper is organized as follows. We discuss related work in section 2. We then describe both our application-driven and data-driven approaches in sections 3 and 4, respectively. Finally, section 5 concludes the paper and provides a brief outlook on future research.

## 2 Related Work

Several authors have already investigated the possibility of using fine-grained electricity consumption data to provide services beyond billing. In particular, there exists a large body of literature dealing with non-intrusive load monitoring (NILM), which aims at determining the contribution of individual appliances to the overall electricity consumption of a household. To this end, NILM approaches leverage measurements of electricity consumption collected at high frequencies – typically varying between 1 Hz and up to multiple kHz – and do not rely on the installation of additional sensors [20]. The possibility to attribute each appliance its actual share of electricity consumption serves as the basis for novel services – such as feedback to motivate a more thrifty use of electricity [8]. In practical settings the accuracy of NILM approaches suffers from the need for high frequency measurements and for time- and data-intensive training procedures [20, 11]. Our work differs substantially from NILM approaches as we do not aim at determining the consumption of individuals appliances but at identifying high-level characteristics of a household – like the number of people living in it. Furthermore, we do not rely on data sampled at high frequency but can cope with a much coarser granularity of few samples per hour or even less.

Other related approaches focus on electricity consumption data recorded at intervals of the order of minutes or hours (typically 15, 30, or 60 minutes). Using such data several authors have for instance analyzed the evolution of consumption patterns over time [3, 1, 19]. In [3], De Silva et al. propose a data mining framework and introduce an incremental learning algorithm that identifies patterns in electricity consumption data. Correlating these patterns over time, the framework aims at predicting future electricity usage on the consumer side – and thus at supporting supply management on the provider side. Similarly, Abreu et al. employ pattern recognition techniques to recognize habitual electricity consumption behavior [1]. Leveraging self-organizing maps, Verdu et al. can recognize consumption patterns that deviate from a "typical" behavior as well as identify new (commercial) customers [19]. These approaches focus on identifying patterns in the available consumption data but do not link them to specific characteristics of the household or

[1] http://www.ucd.ie/issda/data/commissionforenergyregulation/

commercial building causing the patterns to appear. Furthermore, the results presented in the mentioned and related papers are based on small-scale data sets containing traces from a number of household varying between 5 and 30. Instead, in our work we consider consumption traces from more than 3,000 households.

A number of authors have also investigated the problem of clustering consumers into groups that exhibit similar consumption patterns. Knowledge about the existence and characteristics of such clusters can be used to develop novel tariff schemes, improve network management, or to perform load forecasting. An early example of this class of approaches is provided by Chicco et al. [2], who use consumption traces from 471 customers of an electricity provider to perform automatic clustering. Analyzing the resulting clusters and current tariffs of non-residential customers, the authors detect examples of inefficient billing practices (e.g., in case there is a poor correlation between discriminatory factors and actual load patterns) [2].

Several other related approaches that analyze data sets containing traces about a large number of different consumers rely on so-called self-organizing maps (SOMs) [7, 4, 12, 16, 15]. A SOM is an unsupervised learning method based on neural networks that can be used to automatically extract clusters out of an otherwise unstructured (and unlabeled) set of data. For instance, Figueiredo et al. use SOMs to identify groups of consumers with similar consumption behavior [7]. To this end, electricity consumption traces of 165 households are used to train a SOM and accordingly develop a decision tree. In this way, each household can be assigned to a group (i.e., to a cluster in the map) by following a set of rules specified by the structure of the tree. Dent et al. follow a similar approach but base their analysis on a different data set that consists of hourly measurements of electricity consumption of 93 households in the UK [4]. McLoughlin et al. further investigated the problem of automatically clustering consumers with similar consumption patterns [12]. The data set used in this study is however significantly larger than others previously considered and – although this is not explicitly stated in the paper – is most likely to be the same data that we use for our investigation. These three approaches build clusters using only the plain electricity consumption data thus without computing complex features of the data itself. Instead, Sanchez et al. [16] first compute specific features of the data and then feed the SOM with these features – along with additional information obtained through questionnaires. To perform the experiments presented in section 4 we use a number of the features defined in [16]. In contrast to the approaches discussed above, however, the goal of our work is to derive information about the consumers from the electricity consumption data instead of using this data to build classes of consumers that exhibit similar consumption patterns.

Other authors – notably Rasanen et al. – have made use of very coarse-grained consumption data (annual meter readings) – along with detailed information about properties of the corresponding dwellings – to cluster consumers and offer them dedicated energy saving tips [15]. This approach also relies on SOMs but in this case dwelling properties instead of electricity consumption data – or features thereof – are used to train the system. Last but not least, Kolter et al. [10] propose a method that leverages knowledge of dwelling properties (e.g., size, insulation, location) to estimate the most likely energy consumption level of a household within the dwelling. This estimation can the be compared against a household's own actual consumption – as well as against that of similar households. Both Kolter et al.'s and Rasanen et al.'s approaches differ from ours as we aim at estimating household properties based on electricity consumption traces instead of using information about the household to estimate aggregated consumption levels.

## 3 Application-driven Analysis

This section describes the results of interviews that we have performed with employees of three medium-size Swiss energy providers. In the context of these interviews, we focused on a specific consumer-tailored service: energy consulting. The goal of the interviews is to elicit the properties of a household whose knowledge is of highest value to a provider that aims at offering an energy consulting service.

Several energy providers already offer their customers an energy consulting service. However, the service is typically not free and only offered upon request of the customer. Energy providers are however subject to an ever increasing political and social pressure to contribute in reducing the overall energy consumption. As also outlined in our interviews, several providers see the possibility to offer free energy consulting services as a practical way to fulfill the mentioned political "mandate" as well as to please customers. Furthermore, in case of a rise in energy prices the providers also expect the number of customers that request an energy consulting service to increase.

Each interview lasted about two hours and focused on five main topics: 1) Identification and selection of customers representing potential targets of energy consulting services; 2) Typical flow of an energy consulting session; 3) Assessment of potential savings; 4) Determination of potential energy savings; 5) Use of outcome of the energy consulting session (both for the customer as well as the provider); 6) Energy consulting in the long-term.

The analysis of these interviews allows us to make the following qualitative considerations. First, the interviewed energy consultants believe that the availability of additional information about the properties of a household can significantly support the preparation and execution of an energy consulting session. Second, in order to offer energy consulting services on a large-scale and in an efficient way, providers must become able to automatically select customers that are most likely to profit from the service. In the remainder of this section, we discuss these considerations in more detail.

### 3.1 Preparation and Execution of Energy Consulting Sessions

As mentioned above the analysis of our interview data shows that energy consultants consider the availability of information about a household valuable when preparing or executing a consulting session. The respondents indicate as particularly valuable properties like the size of the household

– e.g., expressed in terms of floor area –, the number of bedrooms, and the number of adults and children living in the household. All three respondents stress that even rough estimates of these values, combined with available consumption data, enable the consultant to gain a comprehensive picture of the efficiency of the household and to formulate customer-tailored recommendations. Further properties mentioned by the respondents as particularly valuable include the number and type of electrical appliances present in the household as well as the type of space and water heating.

One of the respondents also notes that hints at potential energy-wasting sources within the household are particularly valuable. Such hints can be obtained by comparing household properties estimated from consumption data with actual properties surveyed with household occupants at the beginning of a consulting session. Interestingly, the respondent also mentioned that about 10% of the private customers served by his company are likely to be affected by wiring errors that might make them pay, for instance, for their neighbor's water heating due to a wrongly connected boiler. The availability of information about household properties makes detection of such flaws much easier for the consultant.

## 3.2 Selection of "High-potential" Customers

In order to make energy consulting services successful, energy providers must become able to identify customers that are likely to be pleased by – and can possibly profit from – such services.

The analysis of our data shows that the interviewed consultants consider two groups of households as particularly interesting targets for an energy consultation: households with a large energy saving potential, and households occupied by certain types of consumers. The first category of households can be identified by looking at high average consumption (e.g., caused by a high number of appliances) as well as the presence of inefficient appliances or of an old infrastructure. The type of heating or cooling used in a household is for instance relevant in order to select "high-potential" customers. As for the second category, one of our respondents points out that retired individuals represent an example of an interesting class of consumers. Indeed, many retired individuals are often ready to invest their time and engage in a consulting session and might be more keen in adapting their consumption behavior. Another of the respondents also indicates "DINK" (Double-Income-No-Kids) households as particularly interesting as their occupants are more likely to invest in renovation measures to improve their energy efficiency.

In summary, our application-driven analysis allows to define the following properties of a household as particularly relevant in order to support energy consulting services: type of employment of the occupants, number of adults/children living in the household, type of space heating, type of water heating, total number of appliances in the household, age of the household dwelling. Whether these characteristics can actually be detected from typical electricity consumption data is discussed in the following section.

# 4 Data-driven Analysis

The goal of the data-driven analysis presented in this section is to provide a list of household properties that can likely be automatically detected through the analysis of electricity consumption data. These properties roughly correspond to the classes that will be included in the first prototype of our household classification system. The data-driven analysis also allows us to verify beforehand the existence of an adequately significant overlap between the properties that are interesting for realistic application scenarios – described in the previous section – and those that can most likely be discovered from the data.

**Table 1. Household properties extracted from the questionnaires accompanying the CER data set.**

| | |
|---|---|
| (1) Properties related to the occupants of the household | |
| Number of adults | `#adults` |
| Number of children | `#children` |
| Number of adults and children | `#persons` |
| Adults at home during day | `#adults@home` |
| Children at home during day | `#children@home` |
| Persons at home during day | `#pers@home` |
| Employment of chief income earner | `employment` |
| Social class of chief income earner | `soc_class` |
| (2) Properties related to the dwelling | |
| Type of dwelling | `type_house` |
| Relationship to property | `own_house` |
| Age of building | `age_house` |
| Floor area | `floor_area` |
| Number of bedrooms | `#bedrooms` |
| (3) Properties related to the appliances in the household | |
| Type of cooking facilities | `type_cook` |
| Type of water heating | `type_water` |
| Type of space heating | `type_heat` |
| Number of appliances | `#appliances` |
| Percentage of energy saving lamps | `lighting` |

Our analysis relies on the publicly available CER data set, a collection of electricity consumption traces from 4,225 private households as well as from 485 non-residential consumers. The traces contain measurements taken at intervals of 30 minutes between July 2009 and December 2010, thus for roughly 1.5 years. The data has been collected as part of the *Smart Metering Electricity Consumer Behaviour Trial (CBT)* carried out between 2008 and 2011 by the Irish Commission for Energy Regulation (CER)[2]. Along with the raw electricity consumption measurements, the data set also contains the answers to questionnaires that have been compiled by the participants of the study before and after the trial. Among other relevant information, these answers reveal a number of interesting properties of the observed households like the number of persons living in the household and their type of employment. Table 1 shows the complete list of properties that we extracted from the questionnaires in order to investigate to what extent they can be inferred from the electricity consumption traces. The experiments presented below are based on a subset of 3,488 out of the 4,225 traces avail-

[2] www.cer.ie

able from private households. The neglected data traces are those for which questionnaires answers were not available.

For carrying out our data-driven analysis we utilize a self-organizing map (SOM) – a well-known method to project high dimensional data onto a 2-dimensional space [9]. A SOM is an artificial neural network that relies on unsupervised learning to group input vectors into *regions* of a map. Each vector is assigned to a specific region depending on its Euclidean (or other type of) distance to already mapped vectors. Clustering procedures can then be applied to group vectors within neighboring regions into *clusters*.

In our analysis, the input vectors consist of features we extract from electricity consumption traces. If a large number of households having the same value for a specific property are mapped to the same cluster on the map, we can conclude that it is also possible to classify households according to this property – or class – using electricity data only. We should however point out that this procedure does not provide a classification of the households. Indeed, we use the SOM to explore the data set and to discover which classes are more meaningful to be included in an automatic classification system. The implementation and evaluation of the classification system itself is part of our future work.

Before discussing the results obtained by analyzing the CER data set using a SOM, we first describe the features we have defined to describe the electricity consumption traces.

## 4.1 Feature Extraction

Building and extending upon related work [16, 2, 7], we define features based on consumption values of individual days as well as aggregated over the entire week or over workdays and weekends separately. In particular, we identify 4 groups of features: (1) consumption figures; (2) ratios; (3) temporal properties; (4) statistical properties.

Consumption figures correspond to simple aggregates of the actual consumption values of a household. For instance, the minimum or maximum consumption values of a day or the average consumption within a specific period (e.g., in the morning or during the night) are referred to as consumption figures. Ratios are quotients of average consumption values of different periods of a day. An example is the ratio between the average consumption in the morning and that during lunch-time. Temporal properties describe the time of the day in which certain events occur. Examples include the time where consumption reaches its daily maximum or the time of the day at which a given consumption threshold is exceeded for the first time. Finally, statistical properties allow to capture qualitative characteristics of the consumption curve. For instance, in order to determine how consumption profiles (of the same household) correlate to each other over subsequent days we compute the cross-correlation between these profiles. Table 2 provides a list of all features we define and use in the context of this work. The table also shows the labels (on the right column) we use to indicate the different features. The intervals *morning*, *noon*, *evening*, and *night* are defined as the time periods 6 a.m. – 10 a.m., 10 a.m. – 2 p.m., 6 p.m. – 10 p.m., and 1 a.m. – 5 a.m., respectively.

To perform our experiments we first compute all the features described above for each household. To this end, we use the consumption data corresponding to a single week of

**Table 2. List of features used to build the input vectors of the self-organizing map. $\bar{P}$ denotes the 30-minute mean power samples provided by the data set.**

| | |
|---|---|
| (1) Consumption figures | |
| $\bar{P}$ (daily) | `c_day` |
| $\bar{P}$ (daily, weekdays only) | `c_weekday` |
| $\bar{P}$ (daily, weekend only) | `c_weekend` |
| $\bar{P}$ for (6 p.m. – 10 p.m.) | `c_evening` |
| $\bar{P}$ for (6 a.m. – 10 a.m.) | `c_morning` |
| $\bar{P}$ for (1 a.m. – 5 a.m.) | `c_night` |
| $\bar{P}$ for (10 a.m. – 2 p.m.) | `c_noon` |
| Maximum of $\bar{P}$ | `c_max` |
| Minimum of $\bar{P}$ | `c_min` |
| (2) Ratios | |
| Mean $\bar{P}$ over maximum $\bar{P}$ | `r_mean/max` |
| Minimum $\bar{P}$ over mean $\bar{P}$ | `r_min/mean` |
| `c_night` / `c_day` | `r_night/day` |
| `c_morning` / `c_noon` | `r_morning/noon` |
| `c_evening` / `c_noon` | `r_evening/noon` |
| (3) Temporal properties | |
| First time $\bar{P} > 1$kW | `t_above_1kw` |
| First time $\bar{P} > 2$kW | `t_above_2kw` |
| First time $\bar{P}$ reaches maximum | `t_daily_max` |
| Period for which $\bar{P} >$ mean | `t_above_mean` |
| (4) Statistical properties | |
| Variance | `s_variance` |
| $\sum(\lvert\bar{P}_t - \bar{P}_{t-1}\rvert)$ for all t | `s_diff` |
| Cross-correlation of subsequent days | `s_x-corr` |
| $\#\bar{P}$ with $(\bar{P}_t - \bar{P}_{t\pm1} > 0.2$ kW) | `s_num_peaks` |

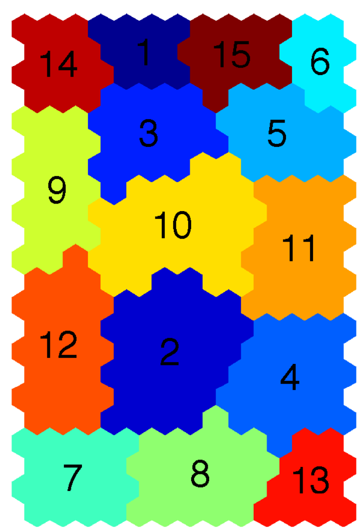
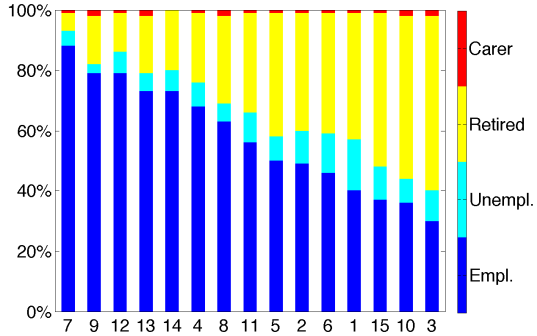
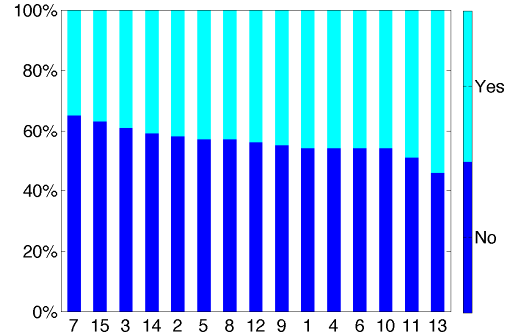


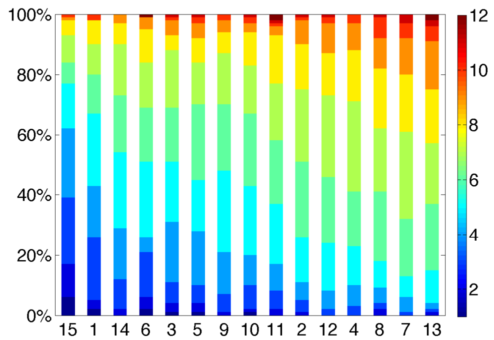

**Figure 1. Clusters obtained when training the SOM using input vectors that contain all the features listed in table 2.**



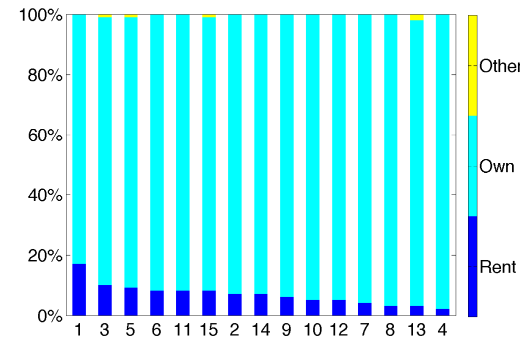

(a) Employment status of the chief income earner of the household (*employment* property).




(b) Time-controlled water heating (*type_water* property).




(c) Number of appliances present in the household (*#appliances* property).




(d) Ownership statistics of the household (*own_house* property).

**Figure 2. Distribution of selected household properties over the different clusters.**

the CER trial. We then normalize the values of the features using the unit variance scaling method [17]. This normalization is necessary since we use the computed features as the components of the input vectors of a SOM. In particular, we use the Euclidean distance as the function that determines the distance between different input vectors and, thus, their position on the SOM. Due to their different nature these features might however exhibit values of very different magnitude. Without normalization, features with large absolute values would thus bias the computation of the Euclidean distance between input vectors and "mask" the effect of features with small(er) magnitudes [17].

## 4.2 Using SOMs to Discover Interesting Household Properties

To implement the SOM we use the *SOM Toolbox 2.0* developed for Matlab by researchers at the Helsinki University of Technology.[3] This tool automatically determines the final number of clusters into which the input data is grouped. In particular, it determines a first set of regions on the map using all the input vectors and then applies a k-clustering filter to balance the map and reduce the total number of regions.

Figure 1 shows the map resulting from feeding the SOM with input vectors that contain all the features listed in table 2 as their components. The input vectors are grouped in 15 clusters. All clusters are of about the same size and contain roughly the same number of households.

While figure 1 displays the final output of the SOM, it does not allow to draw the conclusions we are actually interested in, i.e., if the features computed on the electricity consumption data actually cause households with similar properties to get "naturally" grouped together on the map. To analyze this aspect we can display the percentage of households that exhibit a specific property in each of the clusters identified by the SOM. Figure 2(a), for instance, displays these percentages for the property *employment*, which describes the employment state of the chief income earner (CIE) of the household. The plot shows that the CIEs of about 80% of the households assigned to clusters 7, 9, and 12 are employed, while this percentage decreases to about 30% for clusters 15, 10, and 3. This means that the employment status of the CIE is a property that can likely be discovered from the data, as it is distributed unevenly over the different clusters. In contrast to this, the percentage of households that have a time-controlled water heating is nearly the same in all clusters, as shown in figure 2(b). This means that using the set of features we have defined it is most likely not possible to determine automatically whether a household has a time-controlled water heating or not. Similarly, figure 2(d) also shows that the property *own_house*, which tells whether the occupants of the household are also its owners, can hardly be distinguished according to the clustering provided by the SOM. On the other side, figure 2(c) shows that detecting whether a household has a number of appliances higher than a given threshold might indeed be possible, although not straightforward. In particular, about 80% of the households included in cluster 15 have less than 6 appliances. In con-

[3] www.cis.hut.fi/somtoolbox/

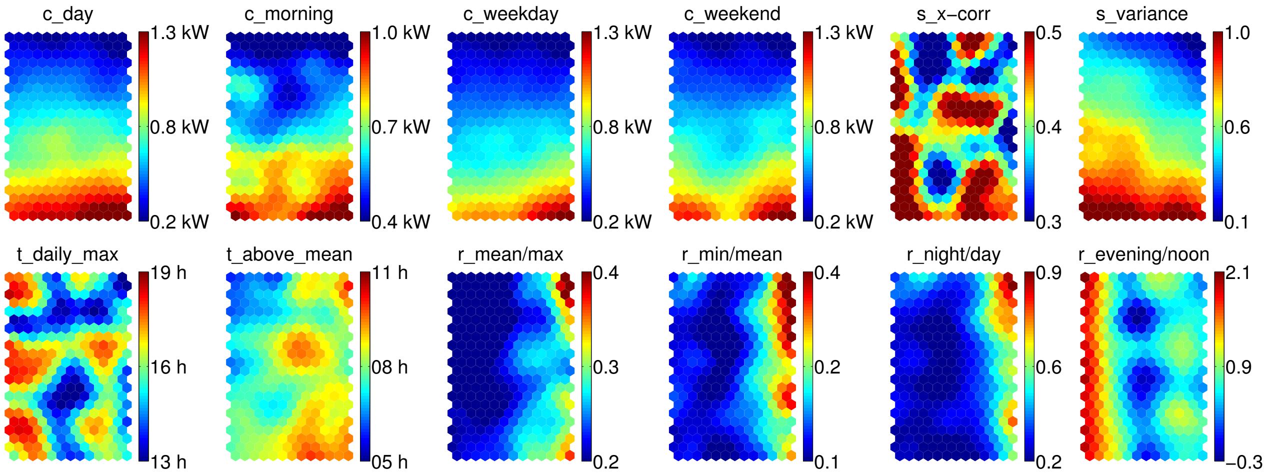



**Figure 3. Component planes of a subset of the features used to train the SOM.**

trast to this, nearly all of the households included in cluster 13 have more than 4 appliances. The results obtained by analyzing the plots relative to all other properties defined in table 1 are not reported here due to space constraints.

As an outcome of our data-driven analysis we can conclude that the following properties are most likely to be inferable from electricity consumption data: *employment*, *type_cook*, *#bedrooms*, *floor_area*, *soc_class*, *#pers@home*, *#persons*, *#appliances*. We thus argue that by applying standard classification techniques to electricity consumption data, it is possible to automatically classify private households according to these properties. The implementation of such a classification system is part of our future work. The other properties listed in table 1 that are not mentioned here appear accordingly less likely to be automatically detectable using electricity consumption data.

## 4.3 Component Planes and Feature Selection

To further analyze the clustering performance of the SOM it is also interesting to observe the so-called *component planes*. A component plane displays how a single feature is distributed over the map or, equivalently, how the feature contributes to the final shape of the map. For instance, the upper left plot in figure 3 shows the component plane relative to the feature *c_day*, which represents the daily electricity consumption of a household. A different color corresponds to a different value of the feature (as indicated by the color bar on the right of each plot). The plot shows that households with a low average consumption are assigned to the top of the map while households with high consumption tend to cluster on the lower right corner. A far less regular distribution is instead exhibited by the component plane relative to the feature *t_daily_max* – also shown in figure 3 – which describes the time of the day at which the maximal value of electricity consumption is reached. Figure 3 displays other examples of features with fairly regular (*c_morning*, *c_weekday*, *c_weekend*, *s_variance*, *r_mean/max*, *r_min/mean*, *r_night/day*, *r_evening/noon*) or quite irregular (*s_x-corr*, *t_above_mean*) component planes.

Features whose component planes exhibit a "regular" distribution over the map are likely to induce a similarly regular structure on the overall map, which is the map that results from a combination of all component planes. In contrast, an "irregular" component plane indicates a feature that does not succeed in inducing a regular clustering. This observation can be used to select a subset of the features that can help improving the final classification results. For instance, training the SOM using only the features with a regular distribution results in a map with only 3 clusters instead of the 15 obtained in the first experiment. Whether this also helps improving the final household classification results will be evaluated in our future work. However, the results obtained observing the component planes of the SOM offer valuable hints to identify the set of features that should first be considered when setting up our envisioned classification system.

## 4.4 Selection of Household Groups

Recalling the results from our interview sessions with the energy consultants, it is also of high interest to energy providers to select "high-potential" households among all customers. Our data-driven investigation shows that properties *employment* and *#appliances* are distributed unevenly over the different clusters of the SOM. We thus further explore our SOM-based analysis to enable the selection of households that very likely have an employed CIE or more than 6 appliances. To this end, we leverage the SOM trained with electricity consumption data as described above. Scanning the SOM we subsequently add households to groups based on their position on the SOM. To determine households whose chief income earner (CIE) is most likely employed, we scan the SOM from the left to the right. This direction is motivated by the fact that clusters that contain households with an employed CIE are located at the left of the SOM depicted in figure 1. Also, this is how features related to consumption ratios (e.g., *r_evening/noon*) – indicating employment – are directed in the component planes. Similarly, we scan the map from the bottom to the top to identify households with a large number of appliances.

To evaluate this household selection method we scan the map two times, adding households to group 1 (by scanning the map from the left to the right) and to group 2 (scanning it from the bottom to the top). We consecutively compute a *success factor* for each group, which we define as the number of households in the group that have an employed CIE (group 1) or more than 6 appliances installed (group 2) – both divided by the total number of households in the group. Figure 4 shows the two success factors depending on the number of households added to the group while scanning the SOM. The x-axis describes the number of households in a group, and the y-axis gives the success factor (i.e., the probability that a randomly selected household from this group has an employed CIE or more than 6 appliances, respectively). The graph shows that it is possible to select a small number of households with a high probability, i.e., of the first 500 households added to group 1, 87% have an employed CIE.

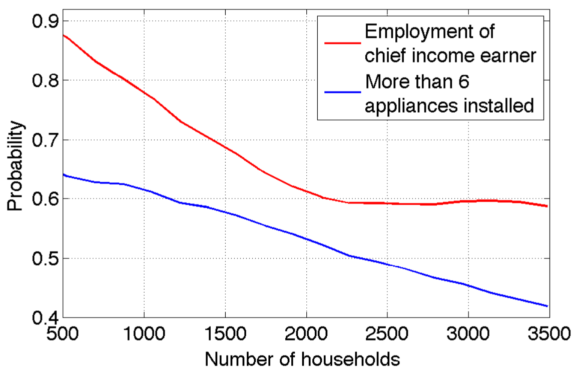



**Figure 4. Success factor of a customer selection method based on self-organizing maps.**

## 5 Conclusions and Outlook

In this paper, we have addressed the problem of performing automatic classification of private households using the data collected by digital electricity meters. In particular, we have identified a set of properties of the households that can be both relevant and promising to look for. To this end, we first elicited a set of relevant properties by performing in-depth interviews with employees of three different energy providers. We then analyzed a large data set of electricity consumption traces using a self-organizing map. Our analysis showed that there exists a set of household properties that are likely to be inferable from electricity consumption data using common classification methods. For instance, our results show that properties like the size of a household and the income of its occupants are both highly useful to energy providers as well as very likely to be inferable from electricity consumption data. The identification of such properties represents a necessary first step towards the investigation of the potential of automatic classification of private households using electricity consumption data. As a next step in this line of research, we will focus on the actual classification problem and address both the selection of adequate classifiers as well as the corresponding optimal feature sets.

## 6 Acknowledgements

The authors would like to thank the interviewees for their valuable contribution. This work has been partially supported by the Hans L. Merkle Foundation, by NCCR-MICS funded by the Swiss National Science Foundation, and by the LOEWE Priority Program Cocoon, a research initiative of the state of Hesse, Germany.

## 7 References

[1] J. M. Abreu, F. P. Câmara, and P. Ferrão. Using pattern recognition to identify habitual behavior in residential electricity consumption. *Energy and Buildings*, 49:479–487, 2012.

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

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

[4] I. Dent, U. Aickelin, and T. Rodden. Application of a clustering framework to UK domestic electricity data. In *11th Annual Workshop on Computational Intelligence (UKCI)*, pages 161–166, 2011.

[5] European Commission. Directive 2006/32/ec of the european parliament and of the council of 5 april 2006 on energy end-use efficiency and energy services and repealing council directive 93/76/eec, 2006.

[6] European Commission. 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, 2009.

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

[8] C. Fischer. Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency*, 1(1):79–104, 2008.

[9] T. Kohonen. *Self-organizing maps*. Springer, 3rd edition, 2001.

[10] J. Kolter and J. Ferreira. A large-scale study on predicting and contextualizing building energy usage. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)*, 2011.

[11] J. Liang, S. K. K. Ng, G. Kendall, and J. W. M. Cheng. Load signature study — part i: Basic concept, structure, and methodology. *IEEE Transactions on Power Delivery*, 25:551–560, 2010.

[12] F. McLoughlin, A. Duffy, and M. Conlon. Analysing domestic electricity smart metering data using self-organising maps. In *CIRED Workshop*, 2012.

[13] K. Metaxiotis, A. Kagiannas, D. Askounis, and J. Psarras. Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. *Energy Conversion and Management*, 44(9):1525–1534, 2003.

[14] F. Nogales, J. Contreras, A. Conejo, and R. Espínola. Forecasting next-day electricity prices by time series models. *IEEE Transactions on Power Systems*, 17(2):342–348, 2002.

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

[16] I. Sánchez, I. Espinós, L. Moreno Sarrion, A. López, and I. Burgos. Clients segmentation according to their domestic energy consumption by the use of self-organizing maps. In *Proceedings of the 6th International Conference on the European Energy Market (EEM 2009), Leuven, Belgium*, pages 1–6, May 2009.

[17] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier Academic Press, 2003.

[18] J. Vasconcelos. Survey of regulatory and technological developments concerning smart metering in the european union electricity market. EUI-RSCAS Working Papers 1, European University Institute (EUI), Robert Schuman Centre of Advanced Studies (RSCAS), Sept. 2008.

[19] S. V. Verdú, M. O. García, C. Senabre, A. G. Marín, and F. J. G. Franco. Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. *IEEE Transactions on Power Systems*, 21:1672–1682, 2006.

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