Customer Classification and Load Profiling Method for Distribution Systems

Antti Mutanen, Maija Ruska, Sami Repo, and Pertti Järventausta

***Abstract*—In Finland, customer class load profiles are used extensively in distribution network calculation. State estimation systems, for example, use the load profiles to estimate the state of the network. Load profiles are also needed to predict future loads in distribution network planning. In general, customer class load profiles are obtained through sampling in load research projects. Currently, in Finland, customer classification is based on the uncertain customer information found in the customer information system. Customer information, such as customer type, heating solution, and tariff, is used to connect the customers with corresponding customer class load profiles. Now that the automatic meter-reading systems are becoming more common, customer classification and load profiling could be done according to actual consumption data. This paper proposes the use of the ISODATA algorithm for customer classification. The proposed customer classification and load profiling method also includes temperature dependency correction and outlier filtering. The method is demonstrated in this paper by studying a set of 660 hourly metered customers.**

***Index Terms*—Clustering, ISODATA, K-means, load profiles, load research.**

I. Introduction

In FINLAND, distribution system loads are commonly estimated with load profiles. Each customer is linked to one of the predefined customer classes, and the load of each customer is then estimated with customer class-specific hourly load profiles. The method involves several error sources and presents significant uncertainties in load estimation. Classification errors are common, because customer classification is based on uncertain customer information. The type of the customer is usually determined through a questionnaire when the electricity connection is contracted. Once the customer type has been determined, it is hardly ever updated. In reality, the customer type may change, for instance, because of a change in the heating solution or an addition of new devices, such as air conditioning. It is a difficult and sometimes impossible task for the system operator to detect the change in customer type only based on billing information. Moreover, the parameters in existing customer class load profiles can be based on measurements, which are old, misclassified , or comprise an insufficient number of measurement points. This is also a significant error source.

Manuscript received August 16, 2010; revised January 27, 2011; accepted April 06, 2011. Date of publication May 19, 2011; date of current version June 24, 2011. Paper no. TPWRD-00616-2010.

A. Mutanen, S. Repo, and P. Järventausta are with the Department of Electrical Energy Engineering, Tampere University of Technology, Tampere FI-33101, Finland (e-mail: antti.mutanen@tut.fi; sami.repo@tut.fi; pertti.jarventausta@tut.fi).

M. Ruska is with the VTT Technical Research Centre of Finland, Espoo FI-02044, Finland (e-mail: Maija.Ruska@vtt.fi).

Digital Object Identifier 10.1109/TPWRD.2011.2142198

Even if the customer information needed in the classification is correct, some of the customers can simply have such an irregular behavior pattern that they do not fit in any of the predefined customer class load profiles. The predefined customer class load profiles also include some inaccuracy due to geographical generalization. The most widespread customer class load profiles are created to model the average Finnish electricity consumption. They do not take into account the regional differences in electricity consumption, which originate from different climate conditions and socioeconomic factors.

Automatic meter reading (AMR) is becoming common in many European countries. AMR provides distribution system operators (DSOs) with accurate and up-to-date electricity consumption data. These data can be used to classify and model distribution network loads. The amount of load data will be enormous when all or almost all of the customers have hourly metering. Since one DSO can have several hundreds of thousands of customers, some kind of automatic data analysis and clustering method should be used.

This paper proposes a pattern-recognition method for customer data classification. The method classifies customers into clusters, for which load profiles can be calculated. These profiles are then used to model customer loads in the distribution system. The method involves temperature dependency correction and outlier filtering.

Different types of clustering techniques have been proposed in the literature for customer classification and load profiling. For example, classical clustering and statistical techniques [1]–[6]; data mining [7], [8]; self-organizing maps [1], [2], [4], [9]; neural networks [10], [11]; and fuzzy logic [4], [5], [10]–[12] have all been applied before.

In previous studies, the customer classification has typically been made according to daily load profiles or load-shape factors. Here, the classification is made according to pattern vectors which include daily, weekly, monthly, and seasonal load variations. Also the motive for customer classification is different. Previously, classification had been studied for the purpose of tariff formulation or marketing strategy planning. Here, the main incentive has been the need for more accurate network calculation: distribution network state estimation [13] and network planning calculation.

The current trend in electricity distribution is to maximize the quality of supply and utilization degree of the existing networks with the help of active network management. Advanced distribution automation functions, such as coordinated voltage and reactive power control as well as automatic feeder reconfiguration and load control require accurate voltage and power-flow estimates. Load model accuracy has a big effect on the distribution network state estimation accuracy [13].

The presented classification method was developed at the VTT Technical Research Centre of Finland. VTT has also developed an application utilizing the presented classification and load profiling method. The LoadModellerPRO program composes load profiles automatically from AMR data and is used by several Finnish distribution system operators. In this paper, the classification and load profiling method are transferred to the MATLAB environment and its classification accuracy is reviewed by comparing it to alternative classification methods.

The presented classification method is universal and can be applied wherever there is sufficient AMR data available. Only the load profiling method needs to be modified to suit local needs and practices. A Finnish case study is presented here. The Finnish distribution system environment provides an excellent platform for the presented method. The hourly load profiles have been in use for a long time, and the AMR installations are increasing rapidly. Finnish DSOs are required to equip at least 80% of the customers with AMR by the end of the year 2013. Section II describes the current Finnish customer classification and load modelling practices. The developed classification method is presented in Section III. Section IV presents some results and Section V discusses the use of the presented classification method. Finally, conclusions are given in Section VI.

II. Load Modeling Method

Finnish load research tradition dates back to the 1980s, when DSOs started to cooperate in load research. The structure of the load model was developed more than 20 years ago. A short description of the Finnish load modeling method is given in Sections II-A-C . In-depth information can be found in [3].

DSOs have customer information systems (CISs), which store all of the available information of each customer's electrical connection, type, and electricity consumption. The customer data usually include:

- electricity connection information: customer location, supply voltage, fuse size, number of phases;
- customer class: residential, agriculture, public, service, industry (NACE code or some other similar code indicating the line of business);
- consumption: annual electricity consumption, high and low tariff electricity consumption (if dual time tariff);
- additional information: heating system (in the case of electric heating: type of electric heating), type of domestic hot water heating system, existence of electric sauna stove.

Traditionally, distribution system estimation uses customer class load profiles for load modeling. Using the information from CIS, each individual customer is linked to one predefined customer class load profile. Finnish DSOs usually use approximately 20–50 customer classes. In addition, some of the largest customers are often modelled with their own models. The customers are also linked to the geographic network model in the network information system (NIS). This enables network calculations using the load profiles.

A. Model Structure

The load model used these days by most Finnish DSOs' software applications represents the expectation value $\mathrm{E}[P(t)]$ and standard deviation $s_P(t)$ for the customer's hourly load as a linear function of the annual energy consumption W_a. The load model can be represented either as topography or as an index series. In topography, the expectation value and standard deviation for hourly load are given for every hour of the year. The expectation value L_{topo} and standard deviation s_{topo} are usually given for a base energy consumption of 10 MWh/yr (W_{base}).

In index series, the load parameters are given in a relative form. The index $Q(t)$ models seasonal variation with 26 two-week indices. The index $q(t)$ models hourly variation for three different day types (working day, Saturday, and Sunday). Each two-week period is modelled separately in index $q(t)$, which thereby consists of $2626 * 3 * 24 = 1872$ indices. Overall, the load expectation values for the whole year are modelled with $1872 + 26 = 1898$ parameters. The hourly standard deviations for the three day types are given as a percentage of the average load in the index $\mathrm{s}_{\%}(t)$.

Formulas for calculating the hourly load parameters (expectation values and standard deviations) with topographies (1) and index series (2) are given

$$\begin{cases} \mathrm{E}[P(t)] = L_{\mathrm{topo}}(t) \cdot W_a / W_{base} \\ s_P(t) = s_{topo}(t) \cdot W_a / W_{\mathrm{base}} \end{cases} \tag{1}$$

$$\begin{cases} \mathrm{E}[P(t)] = \frac{W_a}{8760} \cdot \frac{Q(t)}{100} \cdot \frac{q(t)}{100} \\ s_P(t) = \mathrm{E}[P(t)] \cdot \frac{s_{\%}(t)}{100} \end{cases}. \tag{2}$$

Topographies take special holidays into account, but in the index series, public holidays and eves are modelled as Sundays and Saturdays, respectively. In topographies and in the index series, the reactive power is calculated using one customer class-specific power factor for every hour of the year. In some distribution companies, the reactive power is modeled such as the active power with topographies or index series.

B. Utilization of Load Models

In Finland, loads are modeled down to the individual customer level. Every customer is connected into the network data even at the low-voltage (400 V) level. In distribution network calculation, the customer-level loads are aggregated into higher level loads according to probability theory. For simplicity, loads are assumed normally distributed and independent. In that case, the aggregated load expectation values $\mathrm{E}[P_{ag}(t)]$ and standard deviations $s_{ag}(t)$ for n customers can be calculated with (3) and (4) [3]

$$\mathrm{E}[P_{ag}(t)] = \mathrm{E}[P_1(t)] + \mathrm{E}[P_2(t)] + \ldots + \mathrm{E}[P_n(t)] \tag{3}$$

$$s_{ag}(t) = \sqrt{s_1(t)^2 + s_2(t)^2 + \ldots + s_n(t)^2}. \tag{4}$$

The stochastic nature of the loads is taken into account when calculating peak loads. Load values with different excess probability levels are used in distribution network calculation. The load $P_p(t)$ having an excess probability of $p\%$ can be calculated with

$$P_p(t) = \mathrm{E}[P(t)] + z_p \cdot s_P(t) \tag{5}$$

where z_p is the standard normal deviate corresponding to excess probability p. The load values with excess probability around 10% are relevant for voltage-drop calculation, while smaller probabilities are used when studying loading limits. The load expectation values are used when calculating losses [14].

C. Weather Dependency

The influence of weather on electricity demand is a widely studied phenomenon [15]. Outdoor temperature is usually the single most important factor, but also wind and cloudiness affect electricity demand. In distribution network calculation, a simple weather dependency model is adopted, and only the outdoor temperature dependency is taken into account. In Finland, different electric heating options are widespread, and this, combined with large temperature variations, renders the modeling of the temperature dependency essential in the statistical analysis of customer loads.

As individual loads are metered in different time and location, the effect of temperature variation on a load should be screened out of the data before customer classification. In Finland, a simple and robust model for temperature dependency has been adopted. The temperature-dependent part of the load is modeled as

$$\Delta P(t) = \alpha \cdot (T_{ave} - \mathrm{E}[T(t)]) \cdot \mathrm{E}[P(t)] \tag{6}$$

where

$\Delta P(t)$	outdoor temperature dependent part of the load P at time t;
T_{ave}	average temperature of the previous day;
$\mathrm{E}[T(t)]$	expectation value of the outdoor temperature at time t (long-term daily average temperature);
α	seasonal temperature-dependency parameter [%/°C];
$\mathrm{E}[P(t)]$	expectation value of the load at time t.

In this paper, the parameter α is calculated with linear regression analysis for every four seasons separately. Daily energy consumptions and daily average temperatures are used in the analysis. The effects of daily and monthly fluctuations in electricity demand are eliminated by choosing the regressand and regressor as follows.

- Regressand: the percent error between the daily energy consumption and the average daily energy consumption on a similar day (same weekday and month).
- Regressor: difference between the daily average temperature and the average temperature on a similar day. A one day delay was added to the daily average temperatures to account for the delay in temperature dependency [15].

III. Clustering Method

As the classes and the number of classes are not known beforehand, an unsupervised classification method should be used. In this paper, the iterative self-organizing data-analysis technique (ISODATA) algorithm is used. The algorithm allows the number of clusters to be automatically adjusted if needed.

A. Pattern Vectors

Before the clustering algorithm is applied, each customer's metered load is transformed to a pattern vector. The vector consists of four temperature dependency parameters and 2016 hourly load values. The seasonal temperature dependency parameters are calculated individually for each customer. The achieved parameters are used to normalize the metered load to long-term average temperature. The load values contain weekly average loads calculated for each calendar month.

The annual energies of different customers can vary greatly. The load values in pattern vectors are normalized by dividing each load element by the vector's average load.

B. Outliers

At this stage, outliers are distinguished from other data. Outliers can be failed measurements or customers who use electricity in a very different way from average customers. Two main types of the outliers are as follows.

1) Customers whose electricity use varies significantly during some months. These are detected by comparing each individual customer's monthly energy to the all of the customers' average monthly energy. If a customer's monthly energy differs from the average more than is probable with probability p from the normal distribution, the customer is an outlier. Probabilities between 80% and 99.99% can be applied for this calculation.
2) Customers whose intraday load variation is very high compared to other customers. These customers are filtered out with the help of Euclidean distance measure. The calculation of the Euclidean distance of a pattern vector is described later in Section III-C. If individual customer's Euclidean distance from all customers' average vector is larger than what is probable with probability p from the normal distribution, the customer is classified as an outlier. Probabilities between 80% and 99.99% can be applied for this calculation.

C. Clustering Algorithm

Euclidean distance (7) is chosen for the similarity measure used in the clustering algorithm. The Euclidean distance between two n-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$d_E(x, y) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{7}$$

The first four parameters in the pattern vector are temperature dependency parameters. These parameters are weighted in the analysis. Suitable weights are found experimentally. The weight of the temperature dependency parameters is defined as 5%, and the weight of the actual load measurements is defined as 95%.

The pattern vectors are clustered using the ISODATA algorithm. The method includes heuristic provisions for splitting an existing cluster into two and for merging two existing clusters into a single cluster. The method is unsupervised—the user need not to know the exact number of classes before clustering is completed.

The main procedure of the algorithm is (see, for example, [16] or [17]):

1) Cluster the existing data into c classes but eliminate any data and classes with fewer than T members and decrease c accordingly (Procedure 1). Exit when classification of the samples has not changed.

2) If $c \leq c_d/2$ or $c < 2c_d$ and iteration odd, then
   a) Split any clusters whose samples form sufficiently disjoint groups and increase c accordingly (Procedure 2).
   b) If any clusters have been split, go to step 1.
3) Merge any pair of clusters whose samples are sufficiently close and/or overlapping and decrease c accordingly (Procedure 3).
4) Go to step 1.

Here, c is the number of clusters, c_d is the desired number of clusters, and T is the minimum number of samples in a cluster.

Procedure 1 is a variant of the K-means procedure [18]. A flowchart is shown in Fig. 1.

Procedure 2 for splitting is somewhat heuristic. The flowchart is given later in Fig. 2. ISODATA replaces the original cluster center with two centers displaced slightly in opposite directions along the axis of the largest variance.

The splitting procedure is always performed when the number of clusters is smaller than half the desired number of clusters. Splitting is not performed if the number of clusters is at least twice the desired number of clusters. When the number of clusters is within range $(c_d/2, 2c_d)$, splitting is performed every second round. The desired number of clusters is given by the user and it defines the approximate number of clusters wanted. The final number of clusters also depends on the other user-given parameters and the natural number of clusters in the data.

Two different measures d_k and S_k are used to evaluate the uniformity of the clusters. The quantity d_k is the average distance of samples from the mean of the kth cluster and S_k is the sum of the largest squared distances from the mean along the coordinate axes. Note that here the latter uniformity measure differs from the original (presented in [16] or [17]). Originally, this uniformity was described with a value calculated from only one coordinate axis. In customer load data classification, the uniformity of clusters is better described with information from all coordinate axes

$$d_k = \frac{1}{N_k} \sum_{x \in \chi_k} d_E(x, m_k), \quad k = 1, 2, \ldots, c \tag{8}$$

$$S_k = \frac{1}{N_k} \sum_{i=1}^{n} \max_j \left(x_i^{(j)} - m_{ki} \right)^2, \quad k = 1, 2, \ldots, c \tag{9}$$

where

N_k	number of samples in cluster k;
χ_k	set of vectors belonging to cluster k;
$\boldsymbol{x}_k$	average vector of cluster k;
$d_{\mathrm{E}}(\boldsymbol{x}, \boldsymbol{m}_k)$	distance of vector $\boldsymbol{x}$ from cluster k's average vector;
n	number of elements in the pattern vector;
$x_i^{(j)}$	ith element of pattern vector $\boldsymbol{x}^{(j)}$ belonging to cluster k ($j = 1, 2, \ldots, N_k$);
m_{ki}	ith element of $\boldsymbol{x}_k$.

The overall average distance of samples d is defined by

$$d = \frac{1}{c} \sum_{k=1}^{c} N_k d_k. \tag{10}$$

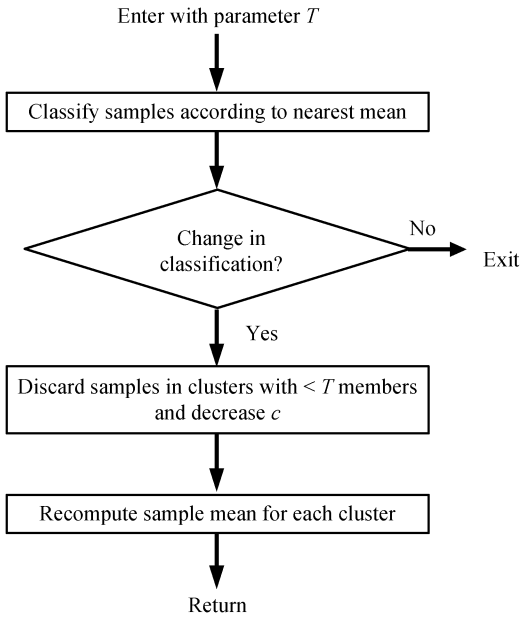


Fig. 1. Flowchart for Procedure 1.

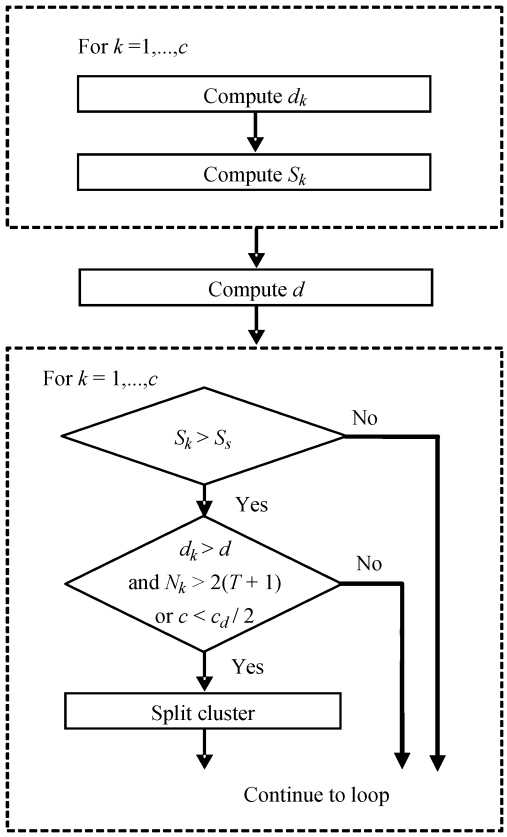


Fig. 2. Flowchart for Procedure 2 (splitting).

The cluster is split if the sum of the largest squared distances from the mean of the cluster $k(S_k)$ is larger than the user-defined threshold value $S_s(S_k > S_s)$ and

$$[d_k > d \text{ and } N_k > 2(T+1)] \text{ or } c < \frac{c_d}{2}. \tag{11}$$

Procedure 3 for merging is performed only if splitting is not executed. Procedure 3 for merging is shown in Fig. 3. At first, all pairwise distances between cluster centers d_{ij} are calculated

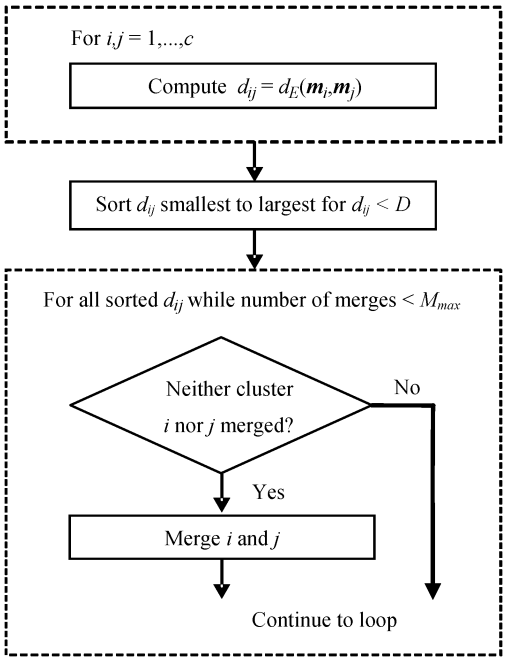


Fig. 3. Flowchart for procedure 3 (merging).

and compared to the threshold value D. Those pairs of clusters corresponding to distances that are less than the threshold value D are arranged in a list from the smallest distance to the largest. The clusters are then merged according to the list's order. Merging continues as long as the total number of merges does not exceed the maximum limit (input parameter $M_{\max}$).

IV. Results

The algorithm shown before was written into a MATLAB program, and its performance is studied here using a set of measurements from 660 hourly measured customers. The measurements have been acquired from a distribution network company in Western Finland. The measurement period used in customer classification and load profiling is from August 18, 2008 to August 17, 2009. The available hourly electricity consumption data had 1-kWh/h measurement resolution. Therefore, only large customers with an annual energy consumption larger than 100 MWh/year are studied. Hourly temperature measurements were also available for the studied network area.

A. Measurement Preprocessing and Outlier Filtering

The measured electricity consumption data can contain errors due to faults in metering or communication. Also, data format changes can cause errors. Typically, these errors are seen as missing values or as errors in the order of magnitude.

In this study, the following preprocessing rules were applied: if the measurement contained a missing data interval longer than five hours or the number of the missing data intervals was larger than five, the measurement was omitted from the data set. Missing parts of the data were estimated by using linear interpolation. If the measured hourly value was clearly the wrong magnitude, the right order of magnitude was estimated by comparing it with the magnitude of the previous hourly value.

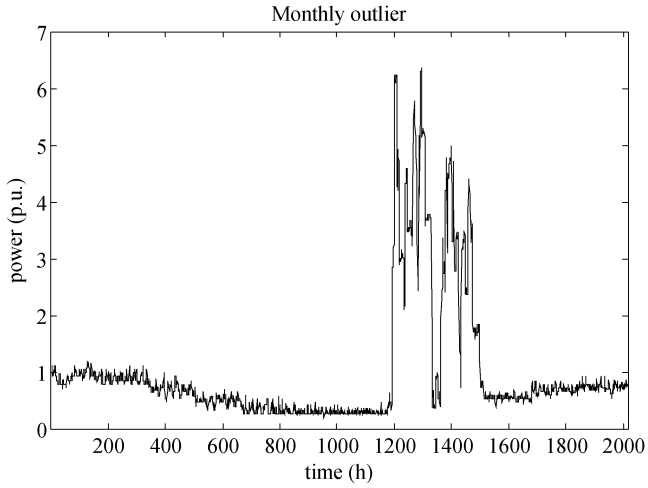


Fig. 4. Pattern vector for a customer with exceptionally large monthly energy consumption in August and September. (Only the load part of the pattern vector is shown.)

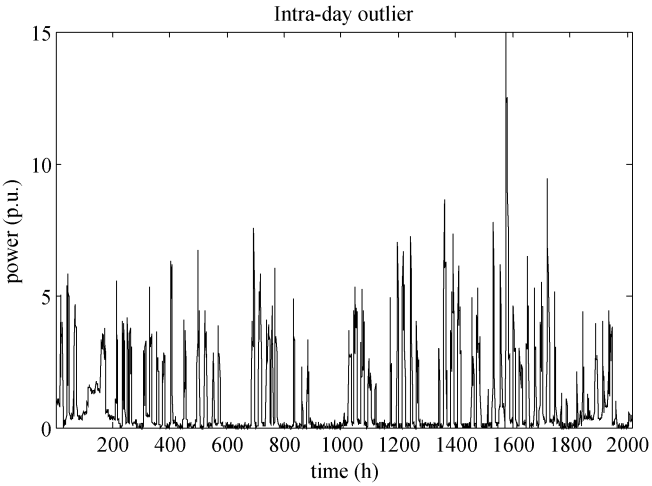


Fig. 5. Pattern vector for a customer with abnormal intraday behavior. (Only the load part of the pattern vector is shown.)

Next, the preprocessed measurements were normalized to long-term (30 years) average temperature, and the individual pattern vectors were formed. Then, the measurements were grouped into six different main customer classes according to the customer class information found in CIS. The selected main customer classes were: residential customers (private apartments and housing corporations combined), agricultural customers, industrial customers, public administration, commercial customers, and other customers (combination of construction, traffic, lighting, and community management).

The outlier filtering was accomplished according to the method presented in Section III-B. A 99% probability level was used to detect abnormalities in monthly energy consumption and a 95% probability level was used to detect abnormal intraday load variations. Examples of the filtered pattern vectors can be seen in Figs. 4 and 5. Note that even if a pattern vector gets filtered, it does not necessarily mean that the corresponding measurement is erroneous; the customer may simply have an extraordinary load pattern.

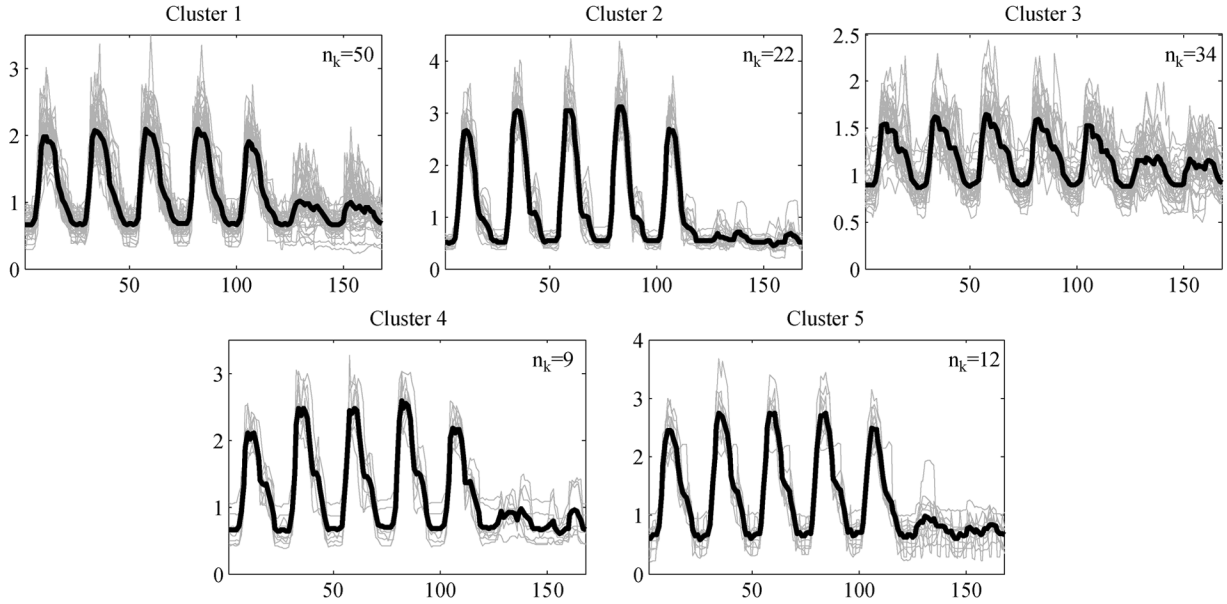


Fig. 6. Results of ISODATA clustering for the public administration main customer class. Horizontal axis: time (h). Vertical axis: normalized load.

The outlier filtering procedure classified 92 out of 660 pattern vectors as outliers.

B. Clustering

The clustering algorithm introduced in Section III-C was used to cluster the remaining 568 pattern vectors. The clustering procedure was carried out separately for each main customer class. Fig. 6 presents the clustering results for the public administration main customer class. For clarity, only the week corresponding consumption in January is presented. The cluster centers are marked with bold black lines and the individual pattern vectors are marked with gray lines. The following parameters were used when clustering public administration customers $c_d = 4$, $T = 1$, $S_s = 30$, $D = 11$, and $M_{\max} = 5$.

The clustering algorithm divided the 127 pattern vectors in the public administration main customer class into five distinct clusters. The number of pattern vectors (n_k) in each cluster varied between 9 and 50. The public administration main customer class contained a total of 151 customers, 24 of them were classified as outliers in the previous step.

Once the classification of the customers is completed, the customer class load profiles can be calculated. The hourly load profiles can be calculated from the original temperature-normalized measurements. The load profiles can be expressed either as topographies or as index series.

Individual load profiles should be used for the outliers. We recommend that the individual load profiles be formed with the same principle as the pattern vectors. That is, the day-type-specific monthly averages are used as expectation values. The use of monthly averages helps smooth out the effect of stochastic variation in the load expectation values. Also, the standard deviations can be calculated when each value is a mean of approximately four hourly values.

The standard deviation calculation is not really reliable if the sample only consists of four hourly values. However, if measurement data are available only from a period of one year, this is a simple way to produce a rough estimate for the standard deviation. After the standard deviations have been calculated, the individual load profiles can be expressed as topographies or index series. In topographies, the average load profile describing one week's consumption is simply duplicated to cover the entire month.

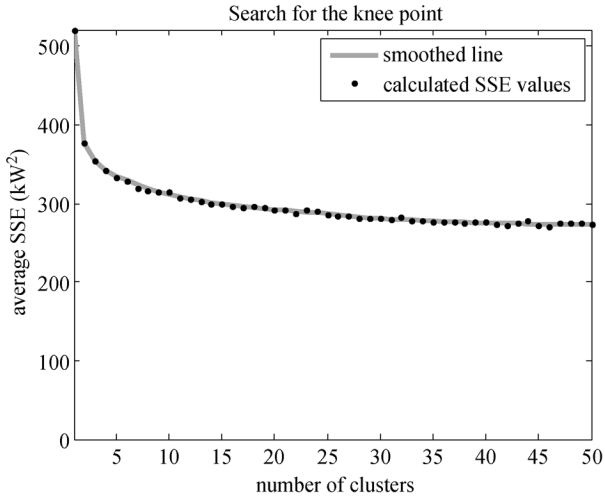


Fig. 7. Public administration load profile SSE as a function of the number of clusters.

The accuracy of load profiles could be increased by increasing the number of customer classes. However, in practice, a compromise between the accuracy and number of customer classes has to be made. Here, the desired number of clusters c_d was selected on the basis of the knee-point criterion [1]. The knee-point criterion helps to find the optimal number of clusters. Fig. 7 shows how the public administration load profile square sum of errors (SSE) between the cluster centers and the measurements depends on the number of the clusters. The knee point is roughly in four clusters. The SSE values in Fig. 7 are calculated similarly as in Section IV-C. For simplicity, the

TABLE I
EFFECT OF THRESHOLD PARAMETERS

parameter		number of actions		consequence
S_s	D	split	merge	
small	small	high	low	large number of clusters
small	large	low	low	bad classification accuracy
large	small	high	high	long computation time
large	large	low	high	small number of clusters

K-means clustering algorithm was used instead of ISODATA when searching for the knee points. In practice, the operator selects the desired number of clusters empirically.

The other user-given parameters also affect the number of clusters. The thresholds for splitting and merging (S_s and D) define how many times the clusters are split and merged. Choosing the right threshold values requires advance information on the type of the customers or use of the trial-and-error technique. High threshold values are chosen when clustering customers with high stochasticity and low thresholds are chosen when clustering customers with low stochasticity. Also, the number of customers affects the threshold values. Table I shows the consequences of choosing too small or too large threshold values. The clustering method is less sensitive to the parameters defining the minimum cluster size (T) and the maximum number of merges (M_{max}). In this study, they were kept in constant values.

C. Accuracy Comparison

To verify the accuracy of the ISODATA clustering method, comparisons were made to alternative classification methods. Classification according to CIS customer class information and allocation to the nearest existing customer class profile were selected as alternative classification methods. The accuracy of the individual load profiles was also verified. The forecasting capability of the load profiles was tested by comparing them with the actual measurements from the time period August 18, 2009 to December 31, 2009. Both expectation and standard deviation values were calculated for the load profiles, but only the load expectation values are studied in these comparisons.

1) Classification Method Comparison: The classification method comparison was made between five different methods: previously presented ISODATA clustering, allocation to the nearest existing customer class profile, and classification in three different accuracy levels according to the CIS customer class information. In this case, the customer class information in CIS is given with a three-digit number. The first number defines the customer's main customer class (e.g. industry), the second specifies classification further (e.g., metal industry), and the third gives the final customer class (e.g., manufacture of metal products). In the level 1 classification, only the first number was used and in levels 2 and 3 also, the second and third numbers were taken into account, respectively. After the classification, CIS-based customer class load profiles were calculated in the same way as the ISODATA-based customer class load profiles. The existing customer class profiles were provided by the Finnish Electricity Association (Sener) [3].

Fig. 8 presents the results for the accuracy comparison. It can be seen that the ISODATA clusters clearly have a smaller square sum of errors than the alternative classification methods, even though some of them had a larger number of customer classes (c). In Figs. 8 and 9, the average SSE is given to measurements normalized to a 10-MWh/yr energy consumption level.

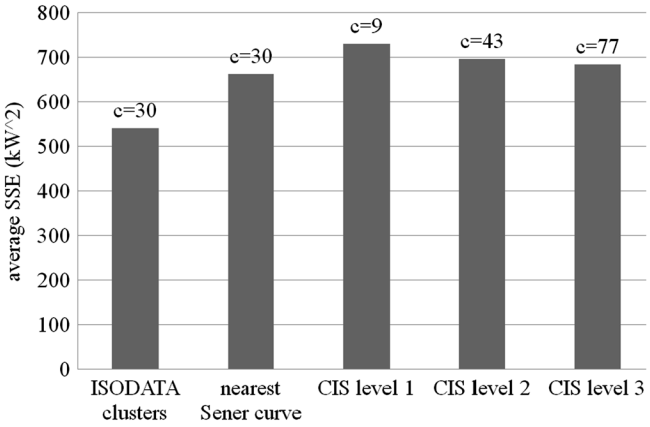


Fig. 8. Comparison of the classification methods.

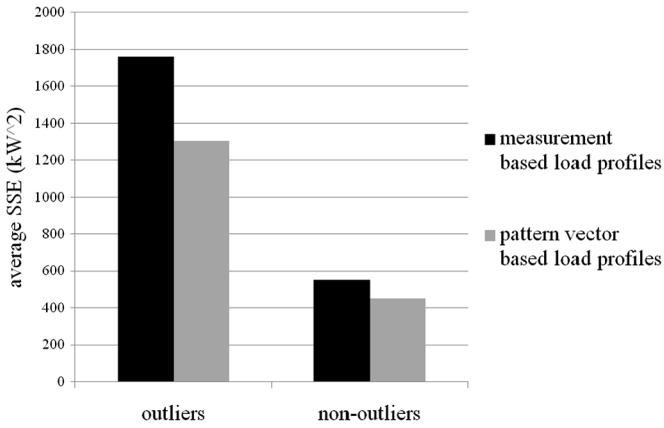


Fig. 9. Individual load profile accuracy comparison.

2) Individual Load Profile Comparison: Here, the individual load profiles were formed based on the pattern vectors. In addition to previous calculations, standard deviations were also calculated for the pattern vector. The original temperature-normalized measurement data were used in the standard deviation calculation. Finally, the load profiles were formed by expanding each section in the pattern vector describing one week's consumption to cover the entire month.

The accuracy of the pattern vector-based individual load profiles was compared to the accuracy of measurement-based individual load profiles. The measurement-based individual load profiles were formed directly from the previous year's measurements corresponding to the studied time period. In individual load profiling, the current practice in distribution companies is to use the previous year's measurements to model the electricity consumption in the current year.

Fig. 9 shows that pattern vector-based load profiles produce better load forecasts than the load profiles formed directly from measurements. Holidays and the temperature dependency were taken into account in both studied load profiling methods. Fig. 9 also shows that load forecasts for the 92 outliers detected in Section IV-A are less accurate than load forecasts for the non-outliers.

V. Discussion

The temperature-dependency calculation, outlier filtering, clustering, and load profile formation for all 660 customers required approximately 60 s of central-processing unit (CPU) time (with a 2.8-GHz Pentium 4 processor), not including the time used for the knee point search. In this paper, all of the customers that passed the outlier filtering were subjected to clustering. In practice, the measurements can be compared with the existing customer class load profiles and only those customers that do not fit the existing load profiles can be subjected to clustering. This can reduce the computation time significantly.

It should be noted that not all the customers should be clustered at the same time. For example, small residential customers should not be clustered simultaneously with large industrial customers. The clustering procedure is based only on expected load values and different-sized customers have different standard deviations. Also, the load model accuracy requirements can be different. Large customers usually have lower stochasticity and, thus, better accuracy can be expected from their load models. Here, this problem was solved by dividing the customers into six main customer classes. However, using the CIS information to divide the customers into the main customer classes can cause new problems. Although rare, it is possible that some customers do not belong to the main customer class specified in CIS. Eliminating this problem would require an additional classification round where the classification of each customer is re-evaluated.

The final number of customer classes depends on how many subtasks that the clustering is divided into and what the desired number of clusters is in each subtask. Ultimately, the operator decides whether he or she wants to emphasize classification accuracy or to keep the number of customer classes easily manageable.

In the study from before, only active power measurements were used. If reactive power measurements are available, the power factors can be taken into account in customer classification and load profiling.

Only customers with a limited amount of missing measurements were used in the clustering. The original measurement set also included measurements with long or frequent periods of missing data. Although the outlier filtering can be used to exclude these failed measurements from clustering, the missing data must be taken into account when forming individual load profiles for outliers. Handling this imperfect measurement series is a challenging task and should be a subject of further research. Also, possibilities to decrease the operator's role in customer classification should be studied.

VI. Conclusions

This paper presents an efficient method for the classification and load profiling of distribution network customers. The classification method utilizes AMR data, is based on the ISODATA algorithm, and involves temperature-dependency correction and outlier filtering. The proposed method was implemented as a MATLAB program and tested with real measurement data. The results showed that the ISODATA algorithm can classify customers into well-separated clusters according to their electricity consumption data. It was also proven that the resulting customer classification is more accurate than the alternative classification methods: classification according to customer class information found in CIS and allocation to the nearest existing customer class profile.

Acknowledgment

The authors would like to thank the Satapirkan Sähkö Oy and its customers for providing the measurement data.

References

[1] G. Chicco, R. Napoli, and F. Piglione, "Comparison among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.

[2] G. Chicco, R. Napoli, and F. Piglione, "Application of clustering algorithms and self organising maps to classify electricity customers," presented at the IEEE PowerTech Conf., Bologna, Italy, Jun. 2003.

[3] A. Seppälä, "Load research and load estimation in electricity distribution," Ph.D. dissertation, Helsinki Univ. Technol., Espoo, Finland, 1996.

[4] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, Aug. 2007.

[5] Z. Zakaria, M. N. Othman, and M. H. Sohod, "Consumer load profiling using fuzzy clustering and statistical approach," in *Proc. 4th Student Conf. Res. Develop.*, Selangor, Malaysia, 2006, pp. 270–274.

[6] I. H. Yu, J. K. Lee, J. M. Ko, and S. I. Kim, "A method for classification of electricity demands using load profile data," in *Proc. 4th Annu. ACIS Int. Conf. Comput. Inf. Sci.*, Jeju Island, South Korea, 2005, pp. 164–168.

[7] B. D. Pitt and D. S. Kirschen, "Application of data mining techniques to load profiling," in *Proc. 21st IEEE Int. Conf. Power Ind. Comput. Appl.*, Santa Clara, CA, 1999, pp. 131–136.

[8] S. Ramos and Z. Vale, "Data mining techniques application in power distribution utilities," presented at the IEEE/Power Energy Soc. Transm. Distrib. Conf. Expo., Chicago, IL, 2008.

[9] S. V. Verdú, M. O. García, F. J. G. Franco, N. Encinas, A. G. Marín, A. Molina, and E. G. Lázaro, "Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters," in *Proc. IEEE Power Eng. Soc. Power System Conf. Expo.*, Atlanta, GA, 2004, pp. 899–906.

[10] K. L. Lo and Z. Zakaria, "Electricity consumer classification using artificial intelligence," in *Proc. 39th Int. Univ. Power Eng. Conf.*, Bristol, U.K., 2004, pp. 443–447.

[11] D. Gerbec, S. Gašperič, I. Šmon, and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," *Proc. Inst. Elect. Eng., Gen., Transm. Distrib.*, vol. 151, pp. 395–400, May 2004.

[12] N. Mahmoudi-Kohan, M. P. Moghaddam, and S. M. Bidaki, "Evaluating performance of WFA K-means and modified follow the leader methods for clustering load curves," presented at the IEEE/Power Energy Soc. Power Syst. Conf. Expo., Seattle, WA, 2009.

[13] A. Mutanen, S. Repo, and P. Järventausta, "AMR in distribution network state estimation," presented at the 8th Nordic Electricity Distribution and Asset Management Conf., Bergen, Norway, Sep. 8–9, 2008.

[14] E. Lakervi and E. J. Holmes, *Electricity Distribution Network Design*, 2nd ed. London, U.K.: Peregrinus Ltd., 1995, ch. 11.3.

[15] M. Meldorf, *Electrical Network Load Monitoring*. Tallinn, Estonia: TUT Press, 2008.

[16] G. H. Ball and D. J. Hall, "ISODATA: A novel method of data analysis and pattern classification," Stanford Res. Inst., Menlo Park, CA, Tech. Rep. NTIS-AD-699616, Apr. 1965.

[17] C. W. Therrien, *Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topic*. New York: Wiley, 1989.

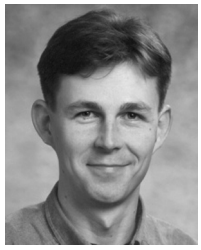
[18] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Symp. Math. Statist. Probability*, Berkeley, CA, 1967, vol. 1, pp. 281–297.

Antti Mutanen was born in Tampere, Finland, on June 10, 1982. He received the M.Sc. degree in electrical engineering from Tampere University of Technology, Tampere, Finland, in 2008.

Currently, he is a Researcher and a Postgraduate student with the Department of Electrical Energy Engineering, Tampere University of Technology. His main research interests are load research and distribution network state estimation.

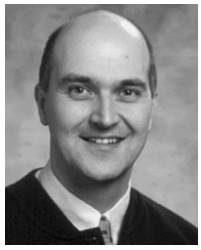
Maija Ruska received the M.Sc. degree in electrical engineering from Helsinki University of Technology, Helsinki, Finland, in 2000.

Currently, she is a Research Scientist at the VTT Technical Research Centre of Finland, Espoo. Her main interests are electricity and fossil fuel markets.



Sami Repo received the M.Sc. and Dr.Tech. degrees in electrical engineering from Tampere University of Technology, Tanin, Finalnd, in 1996 and 2001, respectively.

Currently, he is a University Lecturer with the Department of Electrical Energy Engineering, Tampere University of Technology. His main interest is the management of active distribution networks, including distributed energy resources.



Pertti Järventausta received the M.Sc. and Licentiate of Technology degrees in electrical engineering from Tampere University of Technology, Tampere, Finland, in 1990 and 1992, respectively, and the Dr.Tech. degree in electrical engineering from Lappeenranta University of Technology in 1995.

Currently, he is a Professor in the Department of Electrical Energy Engineering, Tampere University of Technology. His main interests focus on electricity distribution and the electricity market.