Applied Energy 87 (2010) 3538–3545

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data

Teemu Räsänen [a,*], Dimitrios Voukantsis [b,1], Harri Niska [a], Kostas Karatzas [b,1], Mikko Kolehmainen [a]

[a] Department of Environmental Sciences, University of Eastern Finland P.O. Box 1627, FIN-70211 Kuopio, Finland
[b] Department of Mechanical Engineering, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 8 September 2009
Received in revised form 19 January 2010
Accepted 12 May 2010
Available online 8 June 2010

Keywords:
Electricity use
Load curves
Load profiling
Time-series clustering
Self-organizing map
Energy efficiency

ABSTRACT

The recent technological developments monitoring the electricity use of small customers provides with a whole new view to develop electricity distribution systems, customer-specific services and to increase energy efficiency. The analysis of customer load profile and load estimation is an important and popular area of electricity distribution technology and management. In this paper, we present an efficient methodology, based on self-organizing maps (SOM) and clustering methods (K-means and hierarchical clustering), capable of handling large amounts of time-series data in the context of electricity load management research. The proposed methodology was applied on a dataset consisting of hourly measured electricity use data, for 3989 small customers located in Northern-Savo, Finland. Information for the hourly electricity use, for a large numbers of small customers, has been made available only recently. Therefore, this paper presents the first results of making use of these data. The individual customers were classified into user groups based on their electricity use profile. On this basis, new, data-based load curves were calculated for each of these user groups. The new user groups as well as the new-estimated load curves were compared with the existing ones, which were calculated by the electricity company, on the basis of a customer classification scheme and their annual demand for electricity. The index of agreement statistics were used to quantify the agreement between the estimated and observed electricity use. The results indicate that there is a clear improvement when using data-based estimations, while the new-estimated load curves can be utilized directly by existing electricity power systems for more accurate load estimates.

1. Introduction

The analysis of the load profile of customer energy consumption and the estimation of the corresponding load demand are among the priorities of any company related to the production, distribution and management of electricity. There is a constant need to improve the knowledge of load demands in electricity power systems, and this requires for the collection and analysis of load information. On this basis, it is possible to develop better demand models and new customer-oriented applications, that are utilized in tasks such as pricing and tariff planning, distribution network planning and operation, power production planning, load management, customer service and billing and also to provide information to customers and public authorities [1]. These goals are in accordance with the requirements of the European Union legislation [2] concerning energy end-use efficiency and energy services. Furthermore, recent technological progress in the monitoring of the energy use of small customers, is capable of providing with hourly information concerning their load profiles. Such load data have been collected in the past mainly for pricing, quality control and various research purposes [3]. In the future, the availability of such information will allow for more accurate billing and customer-specific services, but, more importantly, will increase energy efficiency due to the better understanding of the energy consumption behavior of customers.

The electricity load curve describes the amount of electric energy a customer uses over the course of time and it is used to plan how much electricity a retailer or distribution company will need to make available at any given time. Furthermore, end-use load curves (i.e., load profiles) reveal the way that customers use electricity at different hours of the day, days of the week and seasons of the year and specify what is customers share to the utility's total load [1]. Additionally, load curves allow for a load estimate in different locations of the distribution networks and provide with better understanding of peak demands [4]. The major factors affecting the customers load profile are (1) customer electricity use behavior and residence characteristics, (2) time of day, week or year and

* Corresponding author. Tel.: +358 44 7162337; fax: +358 17 163191.
E-mail address: teemu.rasanen@uef.fi (T. Räsänen).
[1] Tel./fax: +30 2310 994176.

0306-2619/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apenergy.2010.05.015

(3) local climate factors such as temperature, humidity or solar radiation [1,5].

Typically the energy companies classify customers into groups taking into account their characteristics and annual electricity demand. Based on this classification, each customer is assigned a load estimate curve which is used for billing and energy distribution management. However, it is typical that changes in customer's way of life and electricity use, do not mediate to the energy company in a way that would allow them to update the load curve and thus become more efficient and effective towards market demand. Another problem is that the load curve assigned to a customer may be wrong at the first place due to different electricity use behavior than the proposed typical customer group. Additional problems include the fact that the load curve of different customer groups may possess different statistical characteristics (i.e., follow a different statistical distribution assumption, have differences in the values of basic parameters like the standard deviation, etc.). As a result, the demand side management and distribution planning deals with misinformation causing extra costs.

Modern computational intelligence methods, such as artificial neural networks, support vector machines and self-organizing maps, have been applied in order to forecast energy consumption loads for different prediction horizons in the past [6–9]. Furthermore, recent studies have shown that methods such as K-means, self-organized maps, fuzzy c-means and hierarchical methods, can be applied for the analysis and modeling of the customer electricity consumption behavior [10–12]. In this paper, we present a methodology, which may be applied in complex and large electricity load time-series.

In this study, a large dataset, consisting of 1 year (2008) hourly measured electricity use data for 3989 customers, located in the Northern-Savo, Finland was examined. The hourly electricity use information for large numbers of small customers has been made available only recently. The creation of hourly load curves for small customers has been recently presented, in [13,14]. Nevertheless, hourly data concerning electricity consumption of small customers have not been further analyzed, to the knowledge of the authors. Therefore this paper presents the first results of a detailed analysis of hourly load data, concerning small customers (i.e., customers of a household scale). The aim of the paper was to create a more accurate up-to-date customer-specific load curves using refined measurement data. Furthermore, the purpose was to present efficient way to handle large amount of time-series data and refine this raw data into more valuable information. The resulting load curves were compared with the existing ones for several customer groups, taking also into account the way the load curves are utilized by the company to create load estimates. The results indicate that there is a clear improvement when using the new-estimated, data-based load curves.

2. Materials and methods

2.1. Data used

In order to develop the relevant mean values and standard deviations of load curves for customers belonging to difference classes, large numbers of recorded electricity consumption values are required. Lakervi and Holmes [3] suggest that there must be at least 100 monitored customers for each customer class with electricity use records measured over the last 3 years. Yet, due to the fact that such long datasets are not available At an hourly basis, data used in the current study consisted of 1 year (2008) of hourly measured electricity use data for 3989 customers, located in the Northern-Savo, Finland. The monitoring systems concerning small customers have been established very recently and therefore longer data periods are not yet available. The time-series did not indicate any missing values, since they were selected from a larger database, after passing through a data quality check. An example of a typical electricity use time-series is presented in Fig. 1. The majority of the customers belonged to the Household user group (79.77%), followed by Public Sector (7.92%), Services (6.32%), Agriculture (5.19%) and Industry (0.80%). Typically, the companies further categorize their customers into smaller user sub-groups, depending on information about their house (for household customers) and activities, and assign a predetermined load profile to each one of them. The customers involved in this study had been classified by the company into 18 major user groups.

2.2. Temperature compensation

An important factor that needs to be taken into account during the estimation of load curves is the impact of air temperature on customer's electricity use. Usually the temperature correction of raw electricity use data is performed in order to achieve temperature-free load curves which are not dependent to the temperature variations of the specific year. However, the modeling of the dependence between temperature and electricity use is a difficult task, since relationships tend to vary from customer to customer and pose highly nonlinear nature [15,16]. In Fig. 2, different correlations between customer's electricity use profiles and outdoor air temperature are examplified. It can be clearly seen that detached

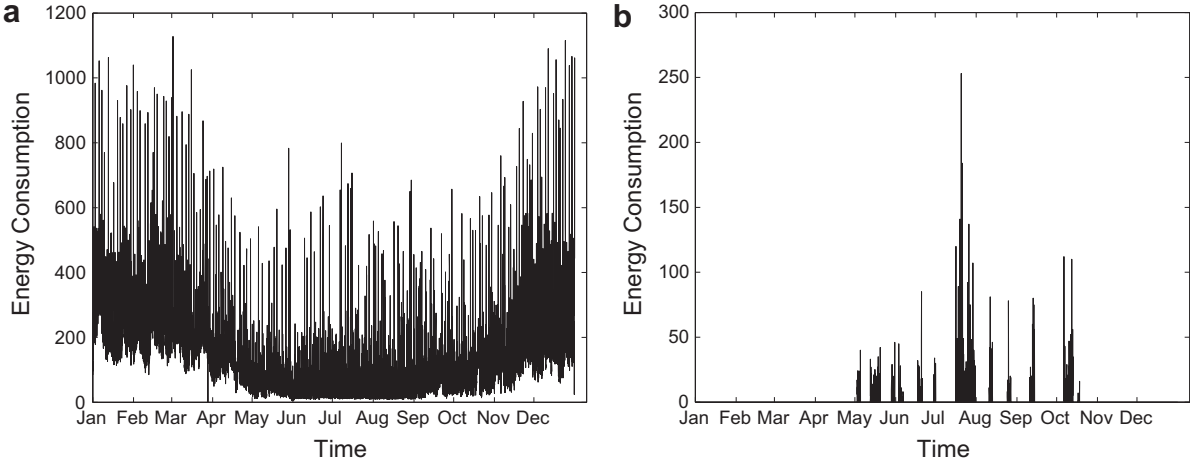

Fig. 1. Typical annual electricity use (kW) for detached house (left) and summer cottage or spare time estate (right).

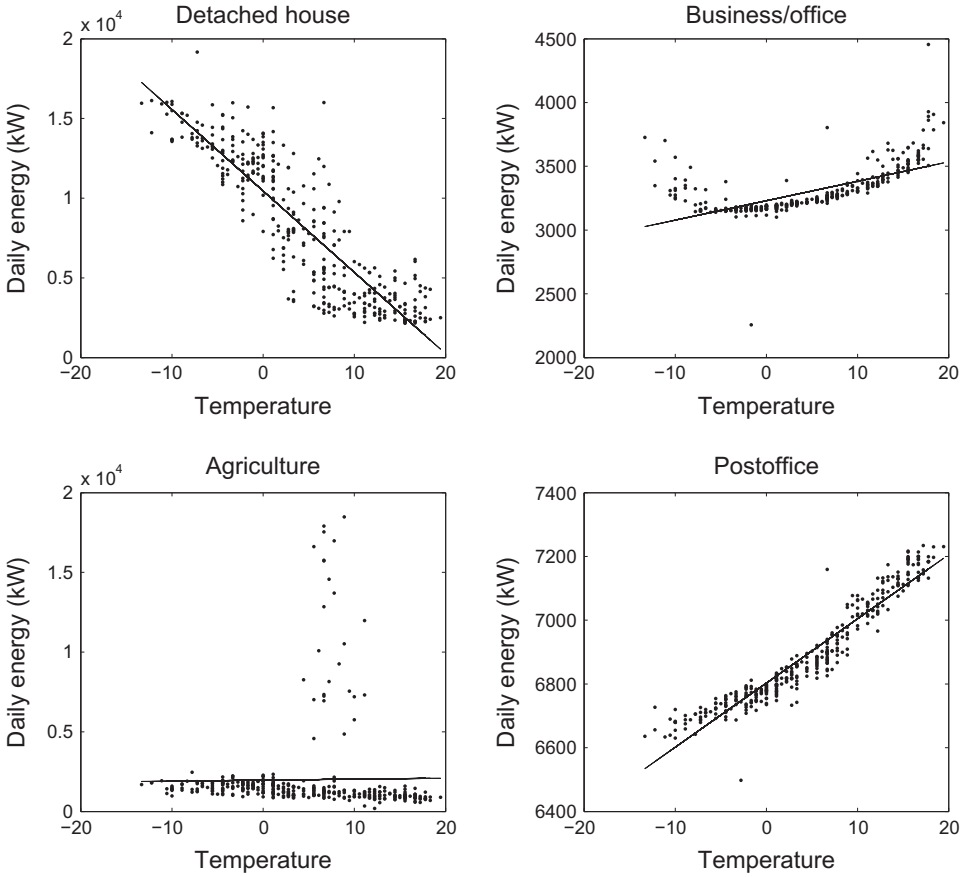

Fig. 2. Four examples of correlation between one customer's daily electricity use and outdoor temperature.

houses use electricity for heating, offices have an air conditioning which is used when outside temperature rises and agricultural customers seem to demonstrate no remarkable correlation between the use of electricity and outdoor air temperature.

In this study, daily average temperature values recorded in the city of Kuopio and were provided by the online weather service of the Savonia University of Applied Sciences [17], and were used to create customer-specific linear regression models for temperature correlation. The fitting was divided into four seasonal patterns (winter, spring, summer and fall) in order to facilitate possible fitting problems caused by seasonal non-linearity. Furthermore, the weekly variation was taken into account by calculating the final temperature correlation factor as the average of the factor defined for workdays, saturdays and sundays in the considered season. Customer-specific indexes [1/°C] were produced to explain a relationship between the percentage of customer's electricity use change [%] and temperature change [1/°C] [25], which were then used to compensate original hourly time-series. Finally, the variance scaling was applied to the compensated time-series in order to allow multivariate clustering analysis of the data.

2.3. Reducing the size of the data

Each customer was represented within the dataset with a yearly time-series of hourly resolution, i.e., 8784 distinct time points. Due to this large amount of data, the application of clustering algorithms would require huge computational resources. Thus, a reduction of the data was required. Typically, there are three different approaches for this task, as described in [18]: (1) raw-data-based approaches, (2) feature-based approaches and (3) model-based approaches. In this case, it was experimentally recognized that there is a clear advantage in terms of performance when the clustering was based on the raw data. Thus, the clustering process was applied to a portion of the initial time-series, corresponding to 5% of the initial length, i.e., 489 time points out of the 8784 initially included in the time-series. The time points were chosen randomly (uniform distribution) and the amount of them was determined experimentally. It has to be noted that these time points were common for all customers so that an inter-comparison is possible, and were used only during the clustering process. This means that during the evaluation process the complete time-series data were used.

2.4. Clustering methods

Although only 5% of the original data, i.e., 489 time points, were used in clustering, the clustering process was a computationally demanding task: electricity use (consumption) data usually contain many peak values. Thus, if clustering is applied directly to the overall time-series, the clustering process may result in a situation where the main characteristic of the customer's behavior is not the main driver for the learning process of the algorithm. Thus, the application of self-organizing maps (SOM) was considered as a suitable intermediate step before the clustering process. Furthermore, the SOM method reduces the size of the data and makes the computational procedure more efficient.

2.4.1. Self-organizing maps

The self-organizing maps method is one of the best known unsupervised neural learning algorithms [19]. The goal of the SOM algorithm is to find prototype vectors that represent the input data set and at the same time realize a continuous mapping from

the input space to a lattice, which is considered to be a mathematical construct topologically representing the "commonalities" between data from the initial data set. SOM is an algorithm characterized by robustness and computational efficiency, thus being an appropriate tool to apply to the large amount of data being made available in this study. Furthermore, it is often combined with other clustering algorithms, serving as an intermediate step of the clustering procedure. There are certain advantages of using this two-step approach that have been already outlined in [20]. Computational efficiency and noise reduction are among the most important ones. In this case, SOM was combined with two different clustering algorithms: (1) K-means clustering and (2) Hierarchical clustering. The lattice of the SOM was chosen to be hexagonal and the map size was 20×20.

2.4.2. K-means clustering

The K-means is a well-known non-hierarchical cluster algorithm [21] that has been applied in many cases in different application domains [22]. In this case, it was combined with SOM in order to make the clustering process computationally efficient and to result in better clustering. The application of K-means requires that the number of clusters is known in advance. Although usually there is no prior knowledge concerning the number of clusters, in this case there are certain restrictions issued by the fact that cluster centers correspond to load curves.

The load models should represent all the customers' classes but deciding the optimal number of classes is a complicated problem. Practically criteria for good quality load data classification are (1) the load variance in one class of customers should be minimized, (2) the number of classes should not to be too large, (3) the classes should be representative and (4) the classes should be easily linked with the energy company's databases [1]. Thus, the choice of the number of clusters has to be balanced between performance and interpretability. Furthermore, a large number of load curves require more effort to be integrated and utilized by the company's system.

Taken these considerations into account a preferred cluster number would be close to the one used by the company, i.e., 18. The exact number was identified by calculating the Davies-Bouldin Index [23] and within cluster variance. Davies-Bouldin (DB) Index is calculated as follows:

$$\mathrm{DB} = \frac{1}{N}\sum_{i=1}^{N}\max_{jj\neq i}\frac{S_i + S_j}{d_{ij}} \tag{1}$$

where N is the number of clusters. The within (S_i) and between (d_{ij}) cluster distances are calculated using the cluster centroids as follows:

$$S_i = \frac{1}{|C_i|}\sum_{x\in C_i}\|x - m_j\| \tag{2}$$

$$d_{ij} = \|m_i - m_j\| \tag{3}$$

where m_i is the centre of cluster C_i, with $|C|$ the number of points belonging to cluster C_i. Additionally, m_j is the centre of cluster C_j. The objective is to find the set of clusters that minimizes Eq. (3).

2.4.3. Hierarchical clustering

SOM was also combined with hierarchical clustering. It is a well-known clustering algorithm [22] that groups several data items together over a variety of scales in order to create a cluster tree. The resulting tree is not a single set of clusters, but rather a multilevel hierarchy of clusters. The similarity of clusters is determined on the basis of a distance metric being applied. In this case, we have used the Euclidean distance of the furthest neighbors, also called complete linkage.

2.5. Estimating the goodness of clustering

The new-estimated load curves were calculated as average values of the electricity use data of the customers belonging to each of the new user groups. The correspondence between customer-specific electricity use and the load curve was presented by calculating the index of agreement (IA).

$$IA = 1 - \frac{\sum_{i=1}^{n}(p_i - a_i)^2}{\sum_{i=1}^{n}(|p_i - \bar{a}| + |a_i - \bar{a}|)^2 \sum_{i}(|p_i - \bar{a}| + |a_i - \bar{a}|)^2} \tag{4}$$

In this case, a_i are the values of each customer's electricity use and p_i are the values of each load curve. Furthermore, a hat is the average of observed data (customer's electricity use-consumption). The IA is a dimensionless measure, limited to the range [0–1], giving a relative size of the difference between an actual (observed) value and its estimation-prediction [24]. It is widely used dimensionless indicator of goodness of model and was not designed to be a measure of correlation but of the degree to which a model's predictions are error free [25]. According to literature the IA is also better suited for model evaluation than, for example, R2, but it is sensitive to extreme values [26]. Values of the IA close to one indicate perfect fit, while values close to zero indicate complete disagreement between the observed and estimated values. Thus, a well formed cluster, with customer of similar electricity use behavior, would be characterized by a high average values and small variation of the IA.

2.6. Comparison between the existing and the new-estimated load curves

The new-estimated load curves were compared to the existing ones for several customer groups as specified by the company. For this purpose we took into account the way the load curves are utilized by the company. Thus, each of the new-estimated load curves was transformed into the so called index series format [27], a file consisting of 26 2-week profiles summarizing hourly electricity use separately for weekdays, saturdays and sundays. Fig. 3 is presenting example of data format for one customer group. In this example we have calculated mean of all 26 weeks for this customer group. Example shows daily electricity use behavior for this group and the difference between weekdays and weekend can be seen easily.

The index series is kind of a dimensionless model which can be scaled using customers annual electricity use in order to calculate estimated electricity time-series for whole year. This procedure is

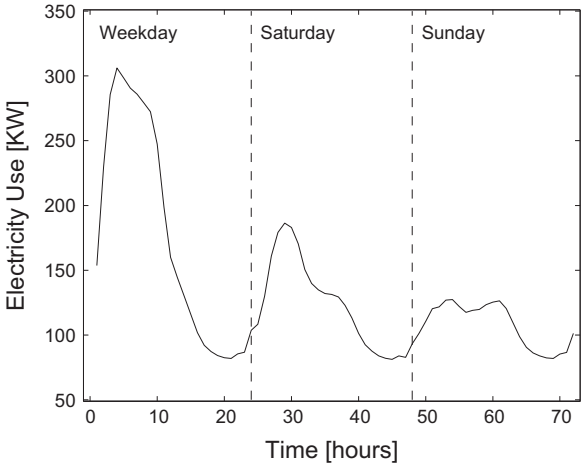

Fig. 3. Mean of 2-week profile for one customer group concerning year 2008.

necessary especially when customer does not have measuring system producing hourly electricity use data. In this way also the peak demand of customers can be estimated. Thus, these index series files are used in order to make hourly, customer-specific estimates of electricity use based on the following equation:

$$P_i = \left(\frac{E}{8736}\right) \times \left(\frac{Q_i}{100}\right) \times \left(\frac{q_i}{100}\right) \quad (5)$$

where P_i is the estimated hourly electricity use, E = the annual electricity use, Q_i = the 2-week electricity use, expressed in percent of the annual average use, and q_i is the hourly electricity use, expressed in percent of the average 2-week use.

3. Results and discussion

3.1. Choosing clustering parameters

The number of time points used for clustering (a subset of 5% of the originally available), was determined experimentally, by evaluating the estimated load curves for several subset sizes. Fig. 4 presents the clustering performance of subsets of different size, and demonstrates the fact that there was a plateau of data points that may be included in the subset in order to estimate the load curves in the most optimum way. The mean index of agreement was calculated for the cluster range 15–30, in order to provide with a reliable estimate of the overall performance. Furthermore, the K-means algorithm was applied five times to avoid a poor performing clustering due to random initialization. Both clustering approaches (SOM + K-means and SOM + Hierarchical), do not indicate any significant improvement if the subset size is larger than the 5% samples used. Furthermore, the SOM + K-means clustering indicates better overall performance than the one indicated by the SOM + Hierarchical clustering.

In most clustering applications the correct number of clusters is not known in advance. In this case, although there was a preferred range for the number of clusters, issued by the company's classification, the exact number was determined by the DB Index and the average within cluster variance. Fig. 5 presents the aforementioned validity indices for several cluster numbers for the SOM + K-means clustering. The DB Index indicates smaller values at 12, 19, 23 and 27 clusters, while the cluster variance is smaller for the second of these choices, i.e., 19 clusters. Taking also into account that the customers involved into this study were classified by the company into 18 major clusters, the choice of 19 clusters was considered for further evaluation, thus allowing a more direct comparison to the classification applied by the company.

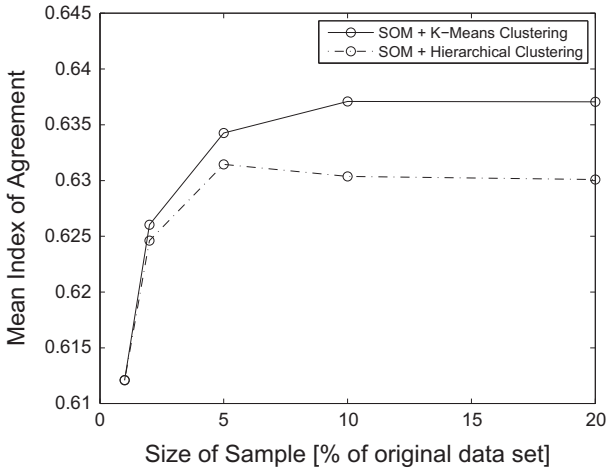

Fig. 4. The performance of the clustering for several subset sizes.

3.2. Evaluation of the new customer groups

The customers involved in this study were divided by the clustering process into 19 distinct user groups, each one characterized by a new-estimated load curve as described in Section 2.4. Table 1 presents the agreement (in terms of the index of agreement) between the new-estimated load curves and the customers identified to belong in the corresponding customer group. Furthermore, the same procedure was applied to the existing customer groups used by the company and the results are presented in the same table for comparison. In Table 1, there are two overall values for mean IA provided. The mean of clusters is calculated using 18 and 19 values provided in Table 1 and the mean of customers is calculated using customer-specific IA values. There is a little difference between these values.

The new-formed customer groups are characterized by higher IA and smaller standard deviation, compared to the clusters resulting from the company's classification. However, there are certain clusters that indicate poor performance, e.g. NLC3. Within this cluster the majorities of customers belong to user groups such as summer cottages or spare time estate. Typically, the electricity use behavior of these user groups is not characterized by certain periodicities or communalities, thus making the behavior of these customers difficult to model. However, these customers usually indicate low electricity use. Thus, although NLC3 includes 8% of the customers involved in this study, these customers correspond only to 2% of the overall electricity use.

The difference between the mean IA of clusters and the overall IA based on customers may be attributed to the formation of small clusters that indicate very high performance and large clusters of poor performance. The IA distribution of all customers is presented in Fig. 6. Although there is still a fraction of the customers that indicate poor performance, there is a clear improvement compared to the clustering applied by the company.

The inspection of customers involved at the new-formed groups, reveals that the clusters include customers from several different domains. Fig. 7 presents an example of that, the customers involved in the groups NLC1 and NLC19 is illustrated. Although clusters might be dominated by certain types of customers (NLC1: Household and NLC19: Public Sector), in all cases there are also customers belonging to other user groups that indicate similar electricity use behavior. This was expected to some extent, since as already mentioned in Section 1, the classification used by the company does not take into account changes in customers' electricity use behavior. Furthermore, the classification applied by the company, may be wrong at the first place, since similar house characteristic or customer activities do not necessarily result in similar electricity use behavior.

3.3. Testing the load curves using estimated annual electricity use

We compare the performance of the new-estimated load curves to the existing ones used by the company in certain customer groups as specified by the company. This validation data, containing 230 customers, was independent and was not used in the clustering. The new-estimated load curves consist of hourly electricity use data and can be utilized directly in order to estimate the load in certain parts of the distribution network. However, they were transformed into the index series format [18], since the existing load curves were available into this format. The index series format is a summarized version of the yearly time-series of 1-h resolution. It separates the 1 year period in 26 2-week periods and the last ones into weekdays, saturdays and sundays. Based on the index

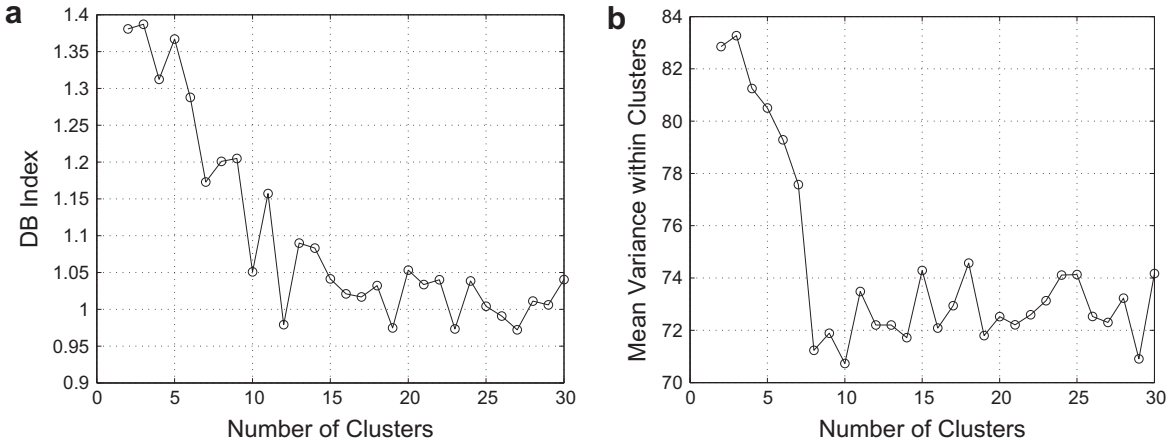

Fig. 5. The DB Index and average within cluster variance for several cluster numbers of the SOM + K-means clustering.

Table 1
The index of agreement for the new customer groups and the ones used by the company.

Company's customer groups				New-estimated customer groups			
Load curve	Mean IA	Std IA	No. of customers	Load curve	Mean IA	Std IA	No. of customers
LC1	0.671	0.097	796	NLC1	0.428	0.072	372
LC2	0.312	0.074	791	NLC2	0.887	0.055	362
LC3	0.316	0.102	379	NLC3	0.130	0.052	317
LC4	0.459	0.119	341	NLC4	0.799	0.044	315
LC5	0.681	0.098	327	NLC5	0.617	0.067	270
LC6	0.135	0.046	314	NLC6	0.672	0.066	265
LC7	0.590	0.090	156	NLC7	0.482	0.070	263
LC8	0.490	0.054	153	NLC8	0.774	0.064	243
LC9	0.562	0.103	133	NLC9	0.589	0.090	239
LC10	0.349	0.107	122	NLC10	0.458	0.089	194
LC11	0.498	0.139	120	NLC11	0.790	0.066	168
LC12	0.578	0.145	84	NLC12	0.302	0.084	160
LC13	0.617	0.160	80	NLC13	0.828	0.058	149
LC14	0.778	0.117	46	NLC14	0.939	0.087	144
LC15	0.941	0.109	36	NLC15	0.605	0.118	141
LC16	0.560	0.182	30	NLC16	0.804	0.083	135
LC17	0.589	0.187	17	NLC17	0.792	0.138	86
LC18	0.720	0.159	11	NLC18	0.722	0.113	84
Other	0.805	0.049	53	NLC19	0.909	0.072	82
Mean of all customers	0.478	0.209	3989		0.627	0.237	3989

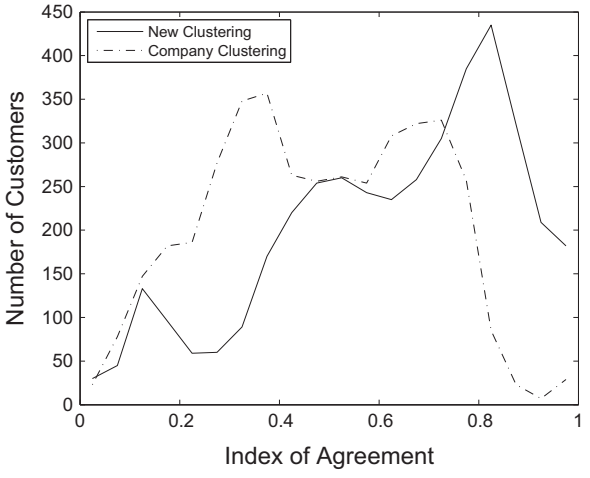

Fig. 6. The distribution of the index of agreement of all the customers involved in the study.

series format a reconstruction of the yearly time-series with 1-h resolution is possible, by using Eq. (5). During this process there is loss of information, however certain periodicities and trends within the time-series are maintained allowing for an estimate of the electricity demand and detection of a peak periods.

In testing the reconstructed time-series data-based on the existing and new-estimated load curves were compared with the actual data, for 230 customers in 10 different groups. The results are presented in Table 2, in terms of the mean IA over all customers belonging in the specific customer group. All customer groups indicate improvement by using the new-estimated load curves. Furthermore, there is two customer groups (detached house, electricity heating, water boiler under 300 L, spare time estate) that indicates clear improvement. Overall, the correspondence between real and estimated (based on load curves) electricity use data was not very high. However, using the presented clustering approach, significant improvements were achieved. In cases such as spare time estates, where the use of electricity is more or less random, was challenging for both approaches. Table 2 includes also standard deviation for both approaches showing that minor improvement was achieved in almost all customers groups when using the new clustering approach.

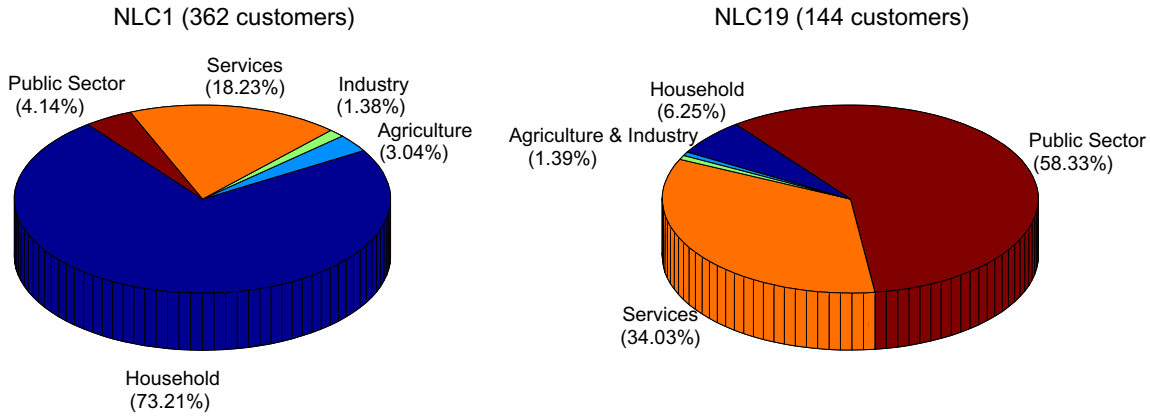

Fig. 7. Customer types involved in 2 of the 19 new formed clusters.

Table 2
Comparison between the actual electricity use data and the estimated ones based on the existing and new-estimated load curves, for five customer groups.

Customer group	No. of customers	IA (existing LC)	Std.	IA (new LC)	Std.
Detached house, electrical heating, water boiler under 300 L (110)	31	0.40	0.10	0.56	0.07
Detached house, partly reserving electrical heating (220)	2	0.35	0.25	0.40	0.13
Detached house, no electrical heating, no electrical stove (601)	19	0.44	0.13	0.56	0.11
Detached house, no electrical heating, electric stove (602)	9	0.55	0.14	0.62	0.07
Block of flats, apartments electricity use included (1020)	6	0.43	0.15	0.49	0.11
Terraced house (whole estate), apartment specific electric heating (1030)	18	0.36	0.08	0.57	0.08
Spare time estate/summer cottage (1140)	5	0.17	0.15	0.39	0.24
Terraced house or block of flats, no electrical heating, no electrical stove (611)	130	0.50	0.21	0.53	0.20
Terraced house or block of flats, no electrical heating, electrical stove (612)	7	0.52	0.14	0.55	0.10
Bank or insurance company (920660)	3	0.32	0.22	0.38	0.25

4. Conclusions

In this paper, we presented a novel computational method capable of handling large amounts of data and applied it for analyzing hourly electricity use data and dividing customers into user groups based on the similarities of their electricity use behavior. The clustering of electricity use time-series data is computationally demanding because of data size and complexity. In this study the clustering was based on raw electricity use data and was applied to a portion of the initial time-series, corresponding to 5% of the initial length. In this way needed computational efficiency was achieved without major effects to the quality of results.

The temperature compensation is a difficult task producing some uncertainty to clustering. In this study, we have presented that the correspondence between outdoor temperature and electricity use is customer-specific and usually nonlinear. Thus, the use of nonlinear modeling methods and a more dense temperature measurement system is one of the main development issues in order to minimize uncertainty in clustering. Moreover, temperature should be measured nearby customer's location, otherwise local conditions cannot be observed.

The presented approach resulted in better estimates of the customers' electricity use (i.e., load curves), compared to the existing ones, originating from classifying customers based on house and activities characteristics. Moreover, it was identified that the new data-based user groups involved customers from several traditional customer groups, thus suggesting that the load curves used for certain customers are not representative for their electricity use behavior. Furthermore, the operational potential of the new-estimated load curves was evaluated using independent data describing 230 customers' electricity use data, initially classified to 10 distinct user groups. It was identified that all customer groups indicated minor or major improvement.

The hourly electricity use for large numbers of small customers has been available only recently. Therefore, this paper presents first results of make major benefits using the data. The results indicate a clear advantage towards data-based estimations that could be utilized directly by existing electricity power systems for more accurate load estimates. Furthermore, the investigation of such detailed information in combination with existing information that has been traditionally utilized by the companies provides a deeper and more complete understanding of customers' demands on electricity.

Acknowledgements

This study was part of ENETE project and scientific collaboration between Research Group of Environmental Informatics (University of Kuopio), Informatics Systems & Applications Group (Aristotle University of Thessaloniki), Savon Voima Verkko Oyj and Enfo Oyj. in order to develop electricity distribution information systems and intelligent services for customers. We would like to thank Mr. Ari Salovaara, Mr. Matti Huovinen, from Savon Voima Verkko Oyj and also Mr. Harri Smolander and Mr. Jouko Kaihua from Enfo Oyj. for providing experimental data, important technical information and guidance during the research project.

References

[1] Seppälä A. Load research and load estimation in electricity distribution. VTT Publications 289, Technical Research Centre of Finland; 1996.
[2] The European Parliament and The Council of the European Union, Directive 2006/32/EC of the European Parliament and of the Council on Energy End-Use Efficiency and Energy Service and Repealing Council Directive 93/76/EEC; 2006.

[3] Lakervi E, Holmes EJ. Electricity distribution network design. 2nd ed. Springer: Berlin; 1995. p. 325. ISBN: 0863413099.
[4] Bartels R, Fiebig DG. Metering and modeling residential end-use electricity load curves. J Forecast 1996;15:415–26.
[5] Elkarmi F. Load research as a tool in electric power system planning, operation, and control – the case of Jordan. Energy Policy 2008;36:1757–63.
[6] Feinberg EA, Genethliou D. Load forecasting. In: Chow JH, Wu FF, Momoh JA, editors. Applied mathematics for restructured electric power systems. Springer; 2005. p. 269–85.
[7] Hippert HS, Pedreira AC, Souza RC. Neural networks for short-term load forecasting. IEEE Trans Power Syst 2001;16:44–55.
[8] Chen B-J, Chang M-W, Lin C-J. Load forecasting using support vector machines: a study on EUNITE competition 2001. IEEE Trans Power Syst 2004;19:1821–30.
[9] Carpinteiro OAS, Reis AJR, da Silva APA. A hierarchical neural model in short-term load forecasting. Appl Soft Comput 2004;4:405–12.
[10] Tsekouras GJ, Kotoulas PB, Tsirekis CD, Dialynas EN, Hatziargyriou ND. A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers. Elec Power Syst Res 2008. doi:10.1016/j.epsr.2008.01.010.
[11] Chicco G, Napoli R, Piglione F. Comparisons among clustering techniques for electricity customer classification. IEEE Trans Power Syst 2006;21:933–40.
[12] Räsänen T, Ruuskanen J, Kolehmainen M. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. Appl Energy 2008;85:830–40.
[13] Hainoun A. Construction of the hourly load curves and detecting the annual peak load of future Syrian electric power demand using bottom-up approach. Int J Electr Power Energy Syst 2009;31:1–12.
[14] Manera M, Marzullo A. Modelling the load curve of aggregate electricity consumption using principal components. Environ Modell Softw 2005;20:1389–400.
[15] Bessec M, Fouquau J. The non-linear link between electricity next term consumption and previous term temperature next term in Europe: a threshold panel approach. Energy Econ 2008;30:2705–21.
[16] Pardo A, Meneu V, Valor E. Temperature and seasonality influences on Spanish electricity load. Energy Econ 2002;24:55–70.
[17] Weather Savonia. Savonia University of Applied Sciences, Online weather service. <http://weather.savonia-amk.fi/>.
[18] Liao W. Clustering of time series data – a survey. Pattern Recogn 2005;38:1857–74.
[19] Kohonen T. Self-organizing maps. 2nd ed. Springer: Berlin; 1997.
[20] Vesanto J, Alhoniemi E. Clustering of the self-organizing map. IEEE Trans Neural Networks 2000;11:586–600.
[21] MacQueen J. Some methods for classification and analysis of multivariate observations. In: Le Cam LM, Neyman J, editors. The fifth Berkeley symposium on mathematical statistics and probability, vol. 1. University of California Press; 1967. p. 281–97.
[22] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv 1999;31:264–323.
[23] Davies D, Bouldin D. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979;2:224–7.
[24] Willmot C. Some comments on the evaluation of model performance. Bull Am Meteorol Soc 1982;63:1309–13.
[25] Harmel RD, Smith PK. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. J Hydrol 2007;337:326–36.
[26] Legates DR, McCabe Jr GJ. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999;35:233–41.
[27] Suomen Sähkölaitosyhdistys ry. Sähkön käytön kuormitustutkimus. Sähköenergialiitto ry.:n julkaisusarja, Helsinki; 1992. ISSN: 0786-7905.