Extracting Controllable Heating Loads from Aggregated Smart Meter Data Using Clustering and Predictive Modelling

Harri Niska

Department of Environmental Science, University of Eastern Finland (UEF)
P.O.Box 1627, 70211 Kuopio, Finland
harri.niska@uef.fi

***Abstract*—Modelling of controllable loads is a necessary function required by demand side management, and specifically load control of smart grids. A large amount of smart metering data and other supporting data are available, enabling the development of new, intelligent data-driven fashions for recognising and modelling loads. However, it is a challenge to extract useful information from this massive, often aggregated data in a reliable and understandable fashion. In this paper we present a data-driven approach, for recognising and modelling of controllable heating loads of small customers. Main computational methods used include self-organizing map (SOM), *k*-means algorithm and support vector regression (SVR). The approach consists of two major stages, namely (i) the recognition of customers that have electrical heating using clustering based on extracted behavioural features and (ii) the predictive regression modelling of controllable heating loads in recognised customer segment. One year of hourly metered electricity consumption data from 525 customers having heterogeneous heating systems, combined with available hourly measured outdoor temperatures and site-specific building information, were used as the base data in the model development and validation.**

I. INTRODUCTION

Today, a massive amount of electricity consumption data is measured worldwide using automatic meter reading (AMR) systems, covering increasingly also small customers. In Finland, for instance, distribution network operators (DSO) are required to install AMR at least 80% of their customers by the end of 2013. AMR data combined with other external datasets such as meteorological observations and building information enable the development of new, intelligent data processing, modelling functions and services required by smart grids.

From the demand side management (DSM) point of view, reliable modelling and prediction of controllable loads and load responses in varying behavioural and environmental conditions is required to ensure sufficient load management and control operations. In Scandinavian countries, the modelling of residential heating loads is of major interest, as they have shown significant control potential compared with other domestic appliances and can be shifted without significantly disturbing the comfort of customers [1].

Until now, a range of sound load modelling approaches can be found in the literature [2−7]. A challenge is, however, that electricity consumption of a customer is recorded in AMR systems as a lump sum rather than being allocated to specific appliances or end uses. This complicates the analysis and modelling of controllable sub loads. Over the past years, techniques for non-instrutive appliance load monitoring (NIALM) are presented, including heating and cooling, lighting, and other domestic appliances [8, 9]. Basically, NILM consists of feature extraction, event detection and pattern recognition. Despite considerable research efforts, reliable load disaggregation is however a challenging task and no complete NIALM solution for all types of household appliances is available [8].

Following the basic principles of NIALM concept, automatic recognition and modelling of controllable loads using the aggregated AMR data could be developed. In recent decades, many advanced data-driven paradigms (data mining, machine learning, evolutionary computation, etc.), contributed by the field of computational intelligence (CI), have been presented. CI methods have shown to yield great capability of processing complex datasets in various application fields, being particularly suitable for:

(i) recognising patterns from high dimensional data
(ii) modelling and forecasting complex time-series influenced by many exogenous variables
(iii) handling measurement errors, noise and incomplete data

The main emphasis of this paper is on the recognition and modelling of controllable electrical heating loads using the AMR data and other supporting information about buildings and outdoor temperatures. In the paper, a novel data-driven CI approach based on a combination of clustering and predictive regression modelling is proposed and evaluated using experimental data. The achieved results are presented and briefly discussed and, finally development ideas and future research outlines are laid out.

II. EXPERIMENTAL DATA

Hourly energy of 525 small customers was measured during the year 2008 using smart meters in the distribution network of Savon Voima Oyj, Finland. In Fig. 1 the measured average powers of the target customers are depicted. In addition to the recorded AMR data, customer-specific building information about primary and secondary heating systems and outdoor temperatures were derived from other available information sources. According to the building information, proportion of heating systems in the target customer group was as follows:

- Oil heating (7%)
- Wood heating (12%)
- Pellet (2%)
- Direct electric heating (22%)
- Direct electric and air source heat pump (6%)
- Ground source heat pump (8%)
- Electric storage heating (3%)
- District heating (40%)

The target data contained thus both electrically and non-electrically heated customers as required by the analysis and modelling. In this study, essential hourly outdoor temperature data were based on the 10 min temperature measurements of representative 15 meteorological stations. The distance weighted average of station-specific hourly temperature measurements was used for estimating prevailing outdoor temperatures in each customer location. In Fig. 2 the average hourly outdoor temperatures measured during the year 2008 are presented.

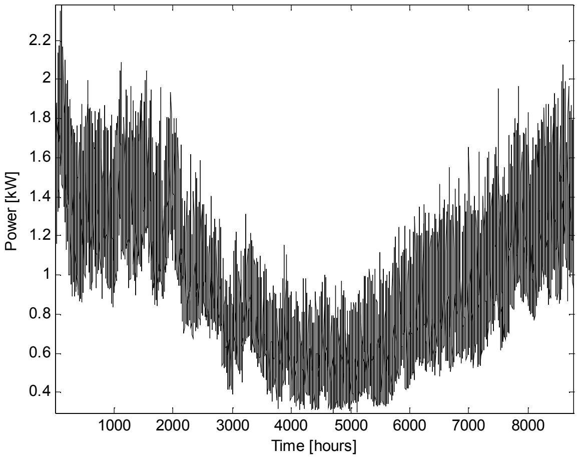

Fig. 1. Average powers of target customers during the year 2008.

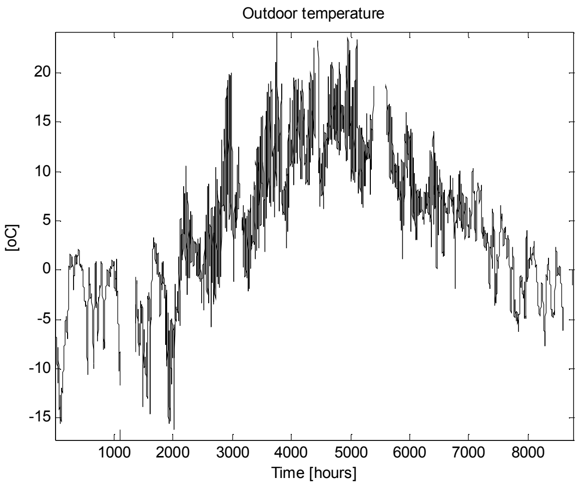

Fig. 2. Average outdoor temperatures during the year 2008.

III. Computational approach and experiments

The approach developed can be divided into two major computational stages, namely: (i) the recognition of electrically heated customers using clustering, and (ii) the predictive modelling of controllable heating loads in the recognised electrically heated segment. Next, these computational stages are described and discussed in more detail.

A. Recognising electrically heated customers

Various clustering methods have been proposed for customer classification and load profiling, including statistical techniques, neural networks, and fuzzy logic [e.g. 10, 11]. The customer classification has been previously made based on daily load profiles, load-shape factors or temporal load patterns [10].

In this study, the clustering was performed based on a subset of features aiming at describe few essential behavioural characteristics, such as temperature dynamics and seasonal, weekly and daily rhythm, of a single customer. The features extracted from the smart meter data were as follows:

- Temperature delay,
  analysed using the linear correlation between the delayed 24h average outdoor temperature and daily energy consumption for each customer.
- Temperature dependence parameters (regression coefficient and constant)
- Daily load profile characteristics (peak hour, minimum load hour, standard deviation, range)
- Daily average energy
- Time series characteristics (curtosis, skewness and variance)

The actual clustering was performed by using self-organizing map (SOM) and the well-known *k*-means algorithm based on the proposed feature variables. SOM is one of the best known unsupervised neural learning methods [12], shown to be particularly suitable for complex data exploration tasks. SOM aims to find prototype vectors that optimally represent the input data, and at the same time to achieve a low dimensional representation of the input space of the training samples usually in the two-dimensional map grid. A particular advantage of SOM is that of ability to speed-up the actual clustering process by reducing the amount of the original smart meter data.

The number of clusters *k* was determined based on visual inspection of formed clusters in respect to available information on customers' heating systems. Visual inspection was performed using Sammon's mapping, which is useful tool for analysis of class distributions and degree of their overlaps in two dimensional space [13]. The basic idea of Sammon's mapping is to represent the points of *p*-dimensional data onto a subspace of two dimensions, preserving the inter-pattern distances as far as possible. However, instead of pure visual inspection as adopted here, it is possible to perform the selection of appropriate number of clusters by means of cluster validity indices such as Davis-Bould index.

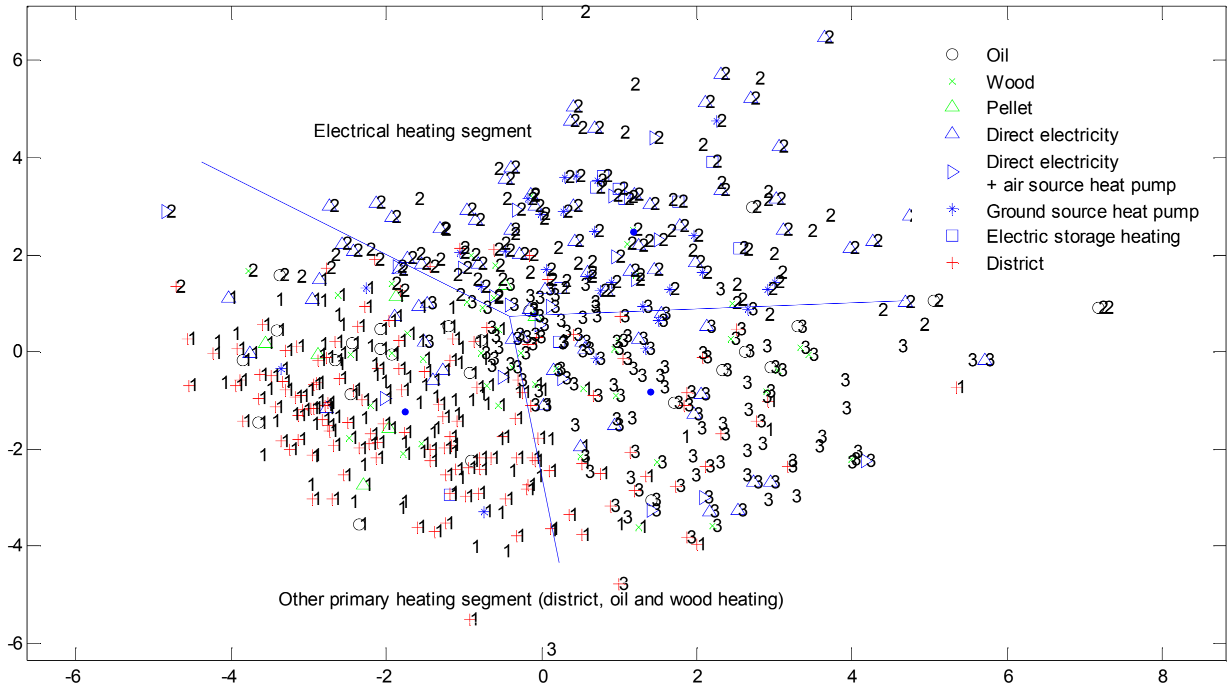

Fig. 3. Sammon's mapping of customers using the extracted feature variables, where the cluster boundaries are visualized based on the cluster centers.

The resulting Sammon's mapping of the customers is presented in Fig. 3. The analysis was observed to be influenced to some extent by errors and inconsistencies in the building information. As a general observation, it can be seen that the calculated feature data are sufficient to discriminate electrically heated customers from non-electrically heated ones with a low number of clusters (here *k*=3). The level of discrimination between the primary heating segments remains still somewhat limited, leaving a room for enhancements in feature extraction.

The recognised electrical heating segment (cluster 2) contains mainly the customers with direct electric heating but also the customers with other electric heating with minor proportion. Reliable discrimination to electric sub segments seems to be limited using the proposed feature data. However, it is possible to perform a rough discrimination into direct electricity and electric storage heating segments using temperature delay as a discriminating factor (here 10 hours was used as a limit value).

In parallel with the clustering, linear discriminant analysis (LDA) was used to classify customers either to electrical heating segment (cluster 2) or other primary heating segment, including district heating, oil heating or wood/pellet heating (clusters 1 and 3). The classification performance achieved was moderately good in terms of performance indices (Kappa index: 0.62, TPR: 87%, and TNR: 82%). The selection of feature variables was performed in respect to Kappa index using multi-objective genetic algorithm [14]. It was found, that 4−6 feature variables out of 13 were capable to produce a sufficient classification performance (Fig. 4). Among the most relevant feature variables were temperature dependence coefficient, temperature delay, daily energy and standard deviation of hourly consumption.

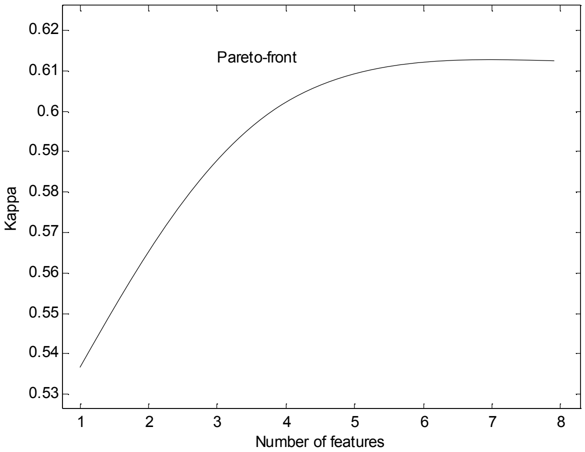

Fig. 4. The observed Pareto-front of explored feature subsets.

B. *Predicting controllable electric heating loads*

Support vector regression (SVR) model was built for predicting hourly average power in the direct electricity segment. SVR based models have emerged relatively new and accurate methods for load modelling and forecasting [15]. Basically, SVR is modern regression method closely related to artificial neural networks [16, 17]. SVR adopts the structure minimization principle, which has been shown to be superior to the empirical risk minimization employed by conventional neural networks.

In this study, commonly used implementation of SVR, namely ε-SVR was adopted [17, 18]. ε-SVR is basically an extension of the linear regression model, which aims to find a function between input variables **x** and target variable **y**.

$$y = \mathbf{w}^T\mathbf{x} + b \qquad (1)$$

where **y** includes the values of the dependent variable (hourly power) and **x** includes the values of independent variables, **w** are the regression coefficients and b includes the residual errors.

The learning task is transformed to the quadratic optimization problem based on the minimization of the so-called Vapnik's ε-insensitive loss function:

$$\min \tfrac{1}{2}\, ||\mathbf{w}||^2 + \mathrm{C}^{n}_{i=1}(\xi_i + \xi_i^*) \qquad (2)$$

Subject to:

$$y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \varepsilon + \xi_i$$
$$\mathbf{w}^T\mathbf{x}_i + b - y_i \leq \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0$$

where n denotes the number of samples, C is a positive constant that defines the trade-off between the training error and the model flatness, ε is the radius of the insensitive zone and ξ are slack variables to measure the deviation of training samples outside ε–intensive zone.

The non-linearity of ε-SVR is achieved by mapping input vector **x** into the higher dimension feature space using a non-linear mapping function, commonly the radial basis kernel function:

$$\exp(-1/2\, \sigma^2\, ||\mathbf{x}_i - \mathbf{x}||^2) \qquad (3)$$

The generalisation power of ε-SVR is highly dependent on control parameters, i.e., the regularisation constant C and the coefficient ε, which control the smoothness of the approximation function and the margin within the error is neglected, respectively. For normalized input signals the value of ε is usually adjusted in the range (10^{-3}–10^{-2}) and C is much bigger than 1.

In this study, ε-SVR (C=10, ε=0.1) with radial basis kernel function (σ=0.1) was employed. The control parameters were selected experimentally. The model output variable y was hourly average power of the target segment and the input variables **x** were day length, hour, divided into continuous sine and cosine components, and outdoor temperature. The selected inputs describe essential factors behind load behaviour, i.e. seasonal and hourly rhythm (household electricity, i.e. appliances and lighting), and temperature dependence (heating). It should be noticed, that the day length variable cannot itself describe day-to-night variation, and thus another weather related variable could be more favourable choice. For the sake of simplicity, weekday was not included here as input variable, although it may have positive influence on performance. On the other hand, it is also possible to build own models for Saturdays and Sundays having different hourly rhythm in terms of household electricity use.

The achieved performance, when assessed by comparing observed and predicted average hourly powers of the segment (Fig. 5), was high in terms of performance indices (R^2=94%, the index of agreement=98% and RMS =0.18). Significant bias in the prediction was not obtained (Fig. 6). The validation was performed here using the hold out scheme, i.e. using 500 random data rows for model fitting, and the rest of the data rows for the validation. The validation between the years was not possible due to the limited data, i.e. the data were limited to the year 2008.

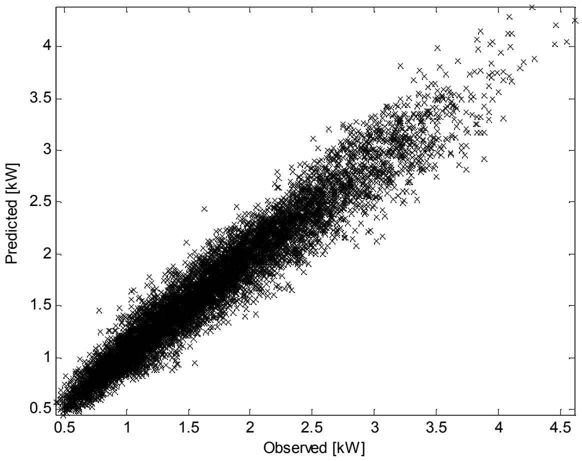

Fig. 5. Observed versus predicted powers of the direct electricity segment.

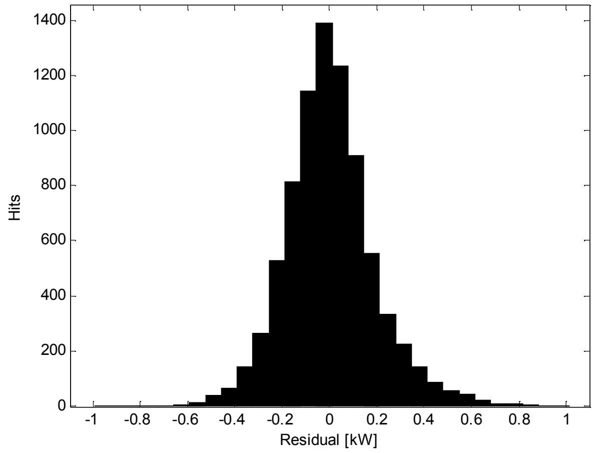

Fig. 6. The histogram of residuals (prediction errors).

The sub loads, i.e. proportion of temperature dependent (controllable heating), hourly dependent and day light dependent loads, were estimated through a simulation. Resulting hourly estimates for the total load and sub loads are presented in Fig. 8. It should be noticed, that the model estimates includes the constant load (~0.5 kW), which can be seen mainly consisting of air conditioning, appliances' standby, etc.

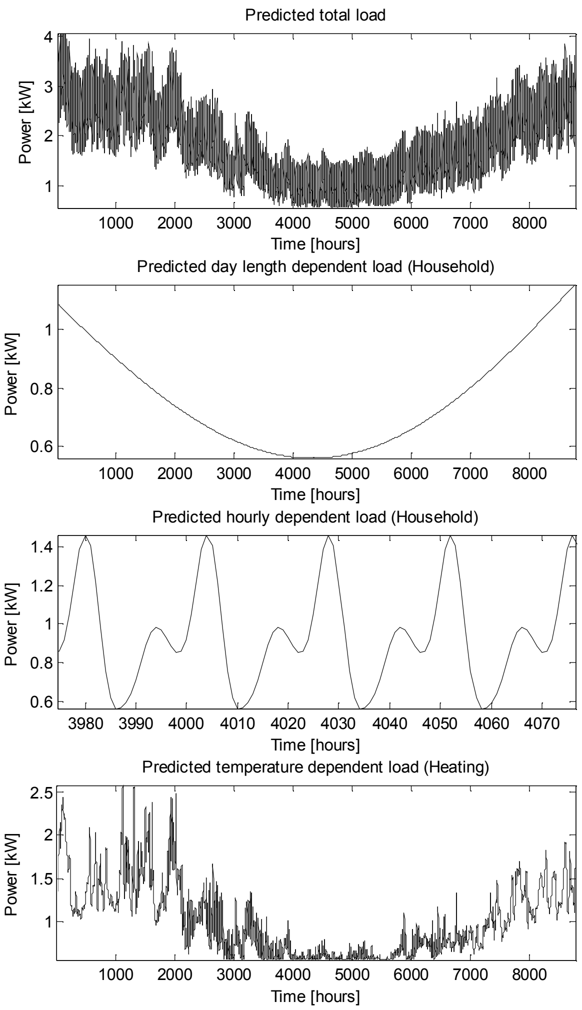

Fig. 8. Predicted total and sub loads of the direct electricity segment (for the sake of clarity, hourly dependent load given for the limited period)

In the case of simulating temperature dependent heating load (Fig. 9), the input vector of the SVR model was constructed by setting hour of day input variable to a value where the consumption of household electricity achieves its minimum (during night time) and the length of day input variable to a value of summer time when the amount of day light achieves its maximum and there is no significant need for lighting. A general challenge of the proposed regression based approach is that the loads are not fully independent, i.e., explanatory variables (outdoor temperature and timing variables) behind the loads correlate to some extent.

The modelling of hourly powers in electric storage heating segment was out of the scope in this study. However, the sub load models extracted from non storage heating segments (such as direct electricity heating and district heating) could be used also in the modelling of these segments.

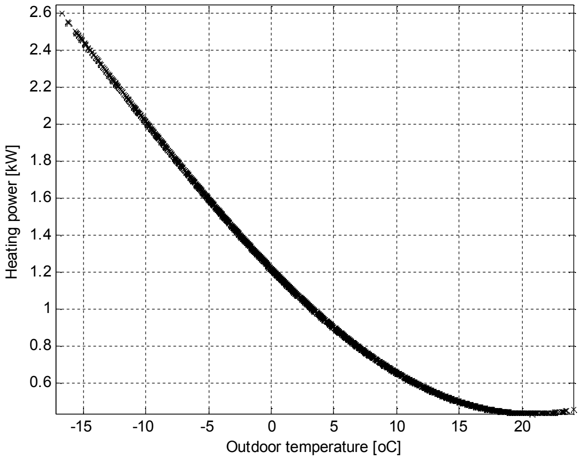

Fig. 9. Estimated temperature dependence of heating power in the direct electricity segment.

IV. Conclusion and Future Work

Reliable modelling of easily managed sub loads is necessary function required by smart grids. In this paper, the data-driven approach, based on the combination of clustering and predictive regression modelling was described and tested for recognising and predicting controllable direct electric heating loads. Various computational methods were applied, among them SOM, *k*-means algorithm, Sammon's mapping and SVR. In general, the results achieved are promising, showing high overall load recognition and modelling performance.

Further developing and testing is however needed in order to enhance of coherence and applicability of the approach. In parallel with the recognition and modelling of primary heating segments and loads, more emphasis should be placed on the modelling of sub loads of electrical heating (e.g. air-to-air heat pumps), as well as electricity storages and micro-generation (e.g. solar panel). In addition, load responses should be taken into account, which may require the development of partly physically based models [3].

Acknowledgment

This work was financially supported by Smart Grids and Energy Market (SGEM) research program of CLEEN Ltd, the Strategic centre for science, technology and innovation of the Finnish energy and environment cluster. The author would like to thank Savon Voima Verkko Oyj and Finnish Meteorological Institute for proving the measurement data, as well as Pekka Koponen (VTT), Göran Koreneff (VTT), Jukka Saarenpää (University of Eastern Finland), Antti Mutanen (Tampere University of Technology), Sami Repo (Tampere University of Technology) and Matti Koivisto (Aalto University) for their valuable comments.

REFERENCES

[1] P. Koponen, “Real-time pricing project at small customers in Finland,” Demand Response Workshop 19 April 2005 in Helsinki.

[2] J. Paatero, and P. Lund, “A model for generating household electricity load profiles,” *International Journal of Energy Research*, vol. 30, pp. 273–290, 2006.

[3] P. Koponen, "Identification of simple physically based models of the response dynamics of electrical heating loads". In Load and response modelling workshop in project SGEM, VTT Working Papers 188, 2011, pp. 39-43.

[4] N. Ruiz, I. Cobelo, and J. Oyarzabal, “A direct load control model for virtual power plant management,” *IEEE Transactions on Power Systems*, vol. 24, pp. 959–966, 2009.

[5] C. Alvarez, A. Gabaldon, and A. Molina, “Assessment and Simulation of the Responsive Demand Potential in End-User Facilities: Application to a University Customer,” *IEEE Transactions on Power Systems*, vol. 19, pp. 1223–1231, 2004.

[6] Molina, A., A. Gabaldon, J.A. Fuentes, and C. Alvarez, “Implementation and assessment of physically based electrical load models: application to direct load control residential programmes,” *IEEE. Proc. Gener. Transm. Distrib.*, vol 150, no 1, 2003.

[7] F. Javed, N. Arshad, F. Wallin, I. Vassileva, and E. Dahlquist, “Forecasting for demand response in smart grids: an analysis on use of anthropologic and structural data and short term multiple loads forecasting,” *Applied Energy*, vol. 96, pp. 150–160, 2012.

[8] M. Zeifman, and K. Roth, “Noninstrutive appliance load monitoring: review and outlook”, *IEEE Transactions on Consumer Electronics*, vol. 57, no. 1, February 2011.

[9] J. Froehlich, E. Larson, S. Gupta, G. Cohn, M. S. Reynolds, and S. N. Patel, “Dissaggregated end-use energy sensing for the smart grid”, *IEEE Pervasive Computing,* vol. 10, no. 1, pp. 28–29, 2011.

[10] A. Mutanen, M Ruska, Repo S., and P. Järventausta, ”Customer classification and load profiling method for distribution systems”, IEEE Transactions on Power Delivery, vol. 26, no 3., July 2011.

[11] T. Räsänen, D. Voukantsis, H. Niska, K. Karatzas, K., and M Kolehmainen, ”Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data,” *Applied Energy*, vol. 87, pp. 3538–3545, 2010.

[12] T. Kohonen, *Self-Organizing Maps*, 3nd, extended edition. Springer, Berlin, 2001.

[13] J.W. Jr. Sammon, “A nonlinear mapping for data structure analysis,” *IEEE Transactions on Computers*, C-18, pp. 401–409, 1969.

[14] N. Srinivas, and K. Deb, “Multiobjective optimization using nondominated sorting in genetic algorithms,” *Evolutionary Computation*, vol. 2, pp. 221–248, 1994.

[15] H. Hahn, S. Meyer-Nieberg, and S. Pickl, “Electric load forecasting methods: tools for decision making,” *European Journal of Operational Research*, vol. 199, pp. 902–907, 2009.

[16] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd Edition. Prentice-Hall, Upper Saddle River, NJ, 1999.

[17] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[18] H. Drucker, C.J.C Burges, L. Kaufman, A. Smola, and V. Vapnik, “Support vector regression machines,” In M. C. Mozer, M. I. Jordan, and T. Petsche (Eds.), *Advances in Neural Information Processing Systems*, volume 9, p. 155. The MIT Press, 1997.