QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 2, 2013, pp. 289 – 297

# FORECASTING OF INDIVIDUAL ELECTRICITY USAGE USING SMART METER DATA

**Tomasz Ząbkowski, Krzysztof Gajowniczek**
Department of Informatics
Warsaw University of Life Sciences – SGGW
e-mail: tomasz_zabkowski@sggw.pl, krzysztof_gajowniczek@sggw.pl

**Abstract:** Forecasting electricity usage is an important task to provide intelligence to the smart gird. The customers will benefit from metering solutions through greater understanding of their own energy consumption and future projections, allowing them to better manage costs of their usage. In this proof of concept paper, we show the approach for short term electricity load forecasting for 24 hours ahead, calculated on the individual household level. In this context authors will develop an approach to the analysis and prediction using Multivariate Adaptive Regression Splines (MARSplines).

**Keywords:** smart metering systems, short term energy forecasting, multivariate adaptive regression splines

## INTRODUCTION

Smart metering is a quite new topic that has grown in importance all over the world and it appears to be a remedy for rising prices of electricity. One of the most important challenge of smart metering is to encourage users to use less electricity through being better informed about their consumption patterns.

Forecasting electricity usage is an important issue to provide intelligence to the smart gird. Accurate forecasting will enable a utility provider to plan the resources and also to take control actions to balance the electricity supply and demand. The customers will benefit from metering solutions through greater understanding of their own energy consumption and future projections, allowing them to better manage costs of their usage.

In this proof of concept paper, our contribution is the approach for short term electricity load forecasting for 24 hours ahead, not on the aggregate but on the

individual household level. The individual customer load profile is influenced by a number of factors, such as devices' operational characteristics, users' behaviours, economic factors, time of the day, day of the week, holidays, weather conditions, geographic patterns and random effects. In this context authors develop an approach to the analysis and prediction of smart metering data using such modelling techniques as Multivariate Adaptive Regression Splines (MARSplines) [Friedman 1991], to capture the factors responsible for accurate short term forecasting in smart metering applications.

Over the last decades different methods have been applied to forecasting the electric load demand. Some of the most popular include time series analyses with autoregressive integrated moving average (ARIMA) [Brockwell and Davis 2002], fuzzy logic [Song et al. 2005], artificial neural network (ANN) [Beccali et al. 2004], [Castillo et al. 2001], [Hippert et al. 2001] and support vector machines (SVM) [Lv et al. 2006]. Majority of them is devoted to analysis of larger loads such as region or the country grid, and therefore, forecasting is achieved with relatively high accuracy [Alfares and Nazeeruddin 2002], [Khotanzad et al. 2002], [Weron 2006].

Leveraging smart metering to support energy efficiency on the individual user level brings research challenges in monitoring usage and providing accurate load forecasting. However, it should be noted that forecasting loads of individual smart meter is not common practice since the volatility of the system is high thus resulting in high error rates [Javed et al. 2012].

## CHARACTERISTICS OF DATA

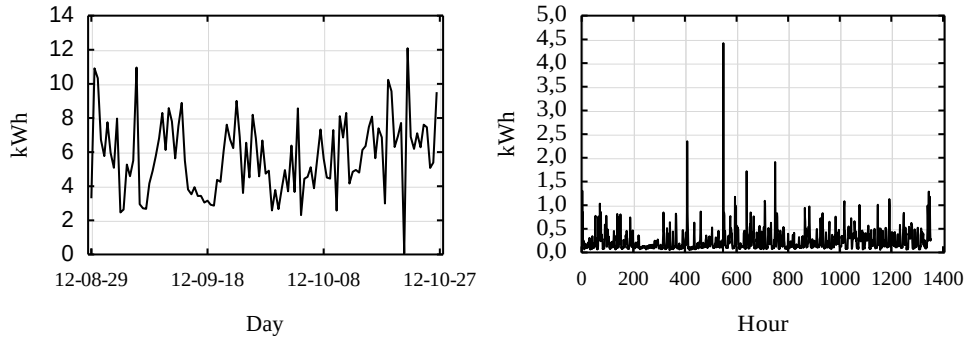
Electricity measurements data were prepared using Mieo HA104 meter installed in one of the households in Warsaw, Poland, for the purpose of SMEPI project[1]. The household consisted of two adult people and a child. The household lived in a flat which was equipped in various home appliances including washing machine, refrigerator, dishwasher, iron, electric oven, two TV sets, audio set, pot, coffee maker, desk lamps, computer, and a couple of light bulbs. The data were gathered during 60 days, starting from 29 August until 27 October 2012.

Original source data contained the electricity usage readings of the Mieosmart meter at every second, every minute and every hour. From these readings, we extracted the hour loads (in kilowatt hour - kWh) for the purpose of short-term load forecasting.

Data characteristics showing the daily and hourly data readings for the analyzed period are illustrated in Figure 1.

[1] SMEPI – Smart Metering Poland, a Hi-Tech project to develop smart metering solutions partially financed by National Centre for Research and Development (NCBiR) and led by Vedia S.A in cooperation with GridPocket and Faculty of Applied Informatics and Mathematics at SGGW.

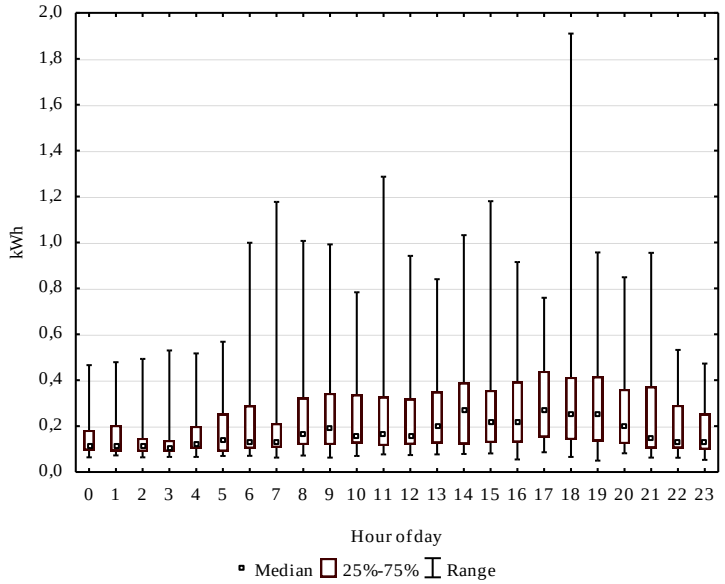
Figure 1. Daily and hourly load in kWh




Source: own preparation

Taking into account that forecasting loads of individual smart meter may be associated with high volatility [Javed et al. 2012] we prepared the box and whisker plot for each of 24 hours using load data over all 60 day, please see Figure 2. The whiskers show the minimum and maximum value in a given hour and box encloses 50% of the total data (top edge represents 75th quartile and bottom edge 25th quartile and line in the middle is the median). The results show that the volatility is rather high (especially during day hours) what can have impact on forecast accuracy.

Figure 2. Box and whisker plot for electricity consumption over each of 24 hours




Source: own preparation

In this research, we focused on forecasting the electricity usage of a particular household for 24 hours ahead. In order to forecast the load we constructed a feature vector with variables as presented in Table 1.

Table 1. Variables used in forecasting

| Variable no. | Description | Formula |
|---|---|---|
| 1 - 24 | Load of previous 24 hours | $W_{hi}, W_{h-1} \ldots W_{h-24}$ |
| 25- 28 | Average load of previous 3, 6, 12, 24 hours | $\frac{1}{i}\sum(W_{hi}),\ i = 3,6,12,24$ |
| 29 - 32 | Maximum load of previous 3, 6, 12, 24 hours | $\max\{W_{hi}\},\ i = 3,6,12,24$ |
| 33 - 36 | Minimum load of previous 3, 6, 12, 24 hours | $\min\{W_{hi}\},\ i = 3,6,12,24$ |
| 37 - 40 | Range of load of previous 3, 6, 12, 24 hours | $\max\{W_{hi}\} - \min\{W_{hi}\},\ i = 3,6,12,24$ |
| 41 | Day of the week | $D_w$ |
| 42 | Day part (morning, noon, afternoon, evening, night) | $D_p$ |
| 43 | Temperature observed in each hour | $T_{hi}$ |

Source: own calculations

These 43 attributes were empirically derived. The individual, the average, the minimum, the maximum and the range loads information were obtained from the hourly load time series. The temperature information inside the flat, for each hour, was collected with Mieo smart meter.

## FORECASTING METHOD

In the experiment we used Multivariate Adaptive Regression Splines (MARSplines) which is nonparametric regression procedure that makes no assumption about the underlying functional relationship between the dependent and independent variables. MARSplines constructs the relation from a set of coefficients and so-called basis functions that are entirely determined from the regression data set. In a sense, the method is based on the divide and conquer strategy, which partitions the input space into regions, each with its own regression equation. In general, nonparametric models are adaptive and very flexible what can ultimately result in over fitting. Although such models can achieve very low error on training data, they have the tendency to perform very bad with new observations. To overcome this problem, MARSplines uses a pruning technique

(similar to pruning in classification trees) to limit the complexity of the model by reducing the number of its basis functions. As basis functions MARSplines uses two-sided truncated functions (hinge function) for linear or nonlinear expansion, which approximates the relationships between the response and predictor variables. A hinge function takes the form:

$$(x-t)_+ = \begin{cases} x-t, & x > t \\ 0, & x \leq t \end{cases} \tag{1}$$

where parameter $t$ is a constant, called the knot, of the basis functions which defining the pieces of the segmented linear regression. It should be stressed, that only positive results of the respective equations are considered, otherwise the respective functions evaluate to zero.

MARSplines can be proposed even in situations where the relationship between the predictors and the dependent variables is non-monotone and difficult to approximate with parametric models, therefore seem much more capable of solving forecasting problem. The general MARSplines model equation is given as

$$P = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) \tag{2}$$

where the summation is over the $M$ nonconstant terms in the model. To summarize, $P$ is predicted as a function of the predictor variables $x$. This function consists of an intercept parameter $\beta_0$ and the weighted by$\beta_m$ sum of one or more basis functions $h_m(x)$.

Implementing MARSplines involves a two-step procedure that is applied successively until a desired model is found. In the first step, the model is build, i.e. increase its complexity by adding basis functions until a preset maximum level of complexity has been reached. After implementing the forward stepwise selection of basis functions, a backward procedure is applied in which the model is pruned by removing those basis functions that are associated with the smallest increase in the goodness of fit.

The so-called Generalized Cross Validation error is a measure of the goodness of fit that takes into account not only the residual error but also the model complexity as well, which is given by:

$$GCV = \frac{\sum_{i=1}^{n}(W_{hi} - P_{hi})^2}{\left(1 - \frac{C}{n}\right)^2} \tag{3}$$

with

$$C = 1 + cd$$

where $W_{hi}$ is the observed load in hour $i$ and $P_{hi}$ is the forecasted load in hour i, n is the number of observations in the original data set, d is the effective degrees of freedom which is equal to the number of independent basis functions and finally,c is the penalty for adding a basis function.

The MARS algorithm can be described and implemented as follows:

1. Start with the simplest model involving only the constant basis function (an intercept parameter $\beta_0$).
2. Search the space of basis functions, for each variable and for all possible knots, and add those which maximize a certain measure of goodness of fit (minimize prediction error).
3. Step 2 is recursively applied until a model of pre-determined maximum complexity is derived.
4. Finally, in the last stage, a pruning procedure is applied where those basis functions are removed that contribute least to the overall (least squares) goodness of fit.

## EVALUATION MEASURES

To assess the model performance for forecasting, we used two measures: precision and accuracy [Javed 2012].

Precision shows how close the model is able to forecast to the actual load. To measure precision we used mean squared error ($MSE$) given by:

$$MSE = \frac{\sum_{i=1}^{n}(W_{hi} - P_{hi})^2}{n} \tag{4}$$

where $W_{hi}$ is the observed load in hour $i$ and $P_{hi}$ is the forecasted load in hour $i$.

In case of accuracy, this measure shows how many correct forecasts the model makes. For this purpose we need to define a correct forecast as the value within a percentage range of the actual load. However, for very low loads, a percentage range may become insignificant. For instance, having a load of 0.1 kWh, a 10% correctness range would be 0.09–0.11 kWh and a forecast of 0.2 kWh will be considered as wrong, but in practice such forecast would be acceptable. To overcome this false loss of accuracy we set two scales to measure accuracy. We set a 10% range of error for accuracy, but if the load is smaller than 1 kWh then we consider range of ±0.10 kWh as range of acceptable forecast. Therefore, accuracy for hour $i$ is given as:

$$\begin{aligned} AC = \sum 1\{W_{hi} > 1 \& |W_{hi} - P_{hi}| < P_{hi} \times 0.10\} \\ + \sum 1\{W_{hi} < 1 \& |W_{hi} - P_{hi}| < 0.10\}. \end{aligned} \tag{5}$$

## ELECTRICITY USAGE FORECASTING

Before estimating and assessing the MARSplines model, we have randomly selected data into two samples. The first set (training) was used to estimate the model, while the second set (test) was used to validate the model. The training and the testing sample included 80% and 20% of the observations, respectively.

As loss function we chose the least squares estimator. In the most general terms, least squares estimation is aimed at minimizing the sum of squared deviations of the observed values for the dependent variables from those forecasted by the model. Technically, the least squares estimator is obtained by minimizing SOS (sum of squares) function:

$$SOS = \sum_{i=1}^{n} (W_{hi} - P_{hi})^2 \tag{6}$$

where $W_{hi}$ is the observed load in hour $i$ and $P_{hi}$ is the forecasted load in hour $i$.

The calculations were prepared in Statistica ver. 10. Due the limitations of the theory and software in the experiment we build 24 models, each for single hour of a day. As a maximum number of basis functions we selected 70 functions, which mean that very complex model will be built in the third step of the algorithm. The forward step usually builds an over fitted model, therefore to build a model with better generalization ability, the backward step prunes the model taking into account c parameter (in our case c=2) which is the penalty for adding a one more basis function. A further constraint which we faced on the forward step is specification of a maximum allowable degree of interaction. Typically, only one or two degrees of interaction are allowed, but higher degrees can be used when the problem require it. Therefore, we choose value 20 as a maximum degree of interaction between independent variables.

The final results obtained by MARSplines and aggregated over all hours are shown in Table 2.

Table 2. Model results aggregated over all hours

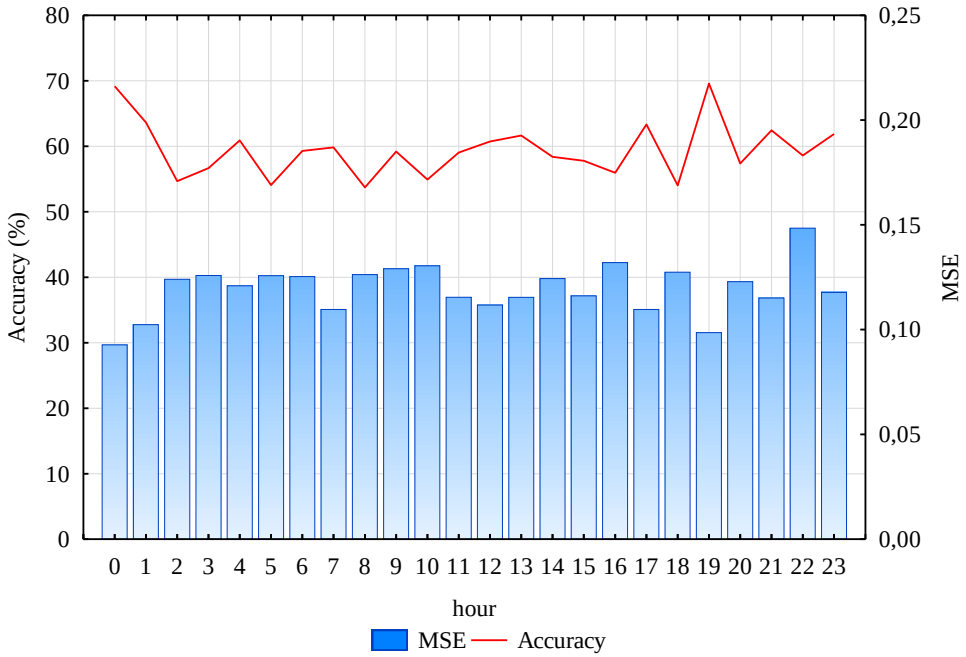
| Measure | Training dataset | Test dataset |
|---|---|---|
| Accuracy (%) | 62% | 59% |
| MSE | 0.10 | 0.11 |

Source: own calculations

For training sample, the accuracy which measures of how many correct forecasts the model makes is 62% and the precision of how close the model is able to forecast to the actual load (MSE) is 0.10. The results associated with the test set are close to these obtained on training set. For this sample MARSplines obtained 59% of accuracy and 0.11 for MSE.

Detailed results per single hour using proposed measures for test sample are shown in Figure 3.

Figure 3. Results in terms of accuracy and MSE for each single hour calculated on the test dataset




Source: own preparation

From the Figure 3 we can observe that almost all hours can be forecasted with relatively high accuracy (more than 50%) which is rather stable over all hours and it does not drops down unexpectedly.

The results presented above are promising but it should be underlined that forecasting on individual household level is perceived as a difficult task since the hourly and daily behaviour may change drastically due to different circumstances, e.g. using particular home appliances, weather conditions, holidays of household members. In larger populations such as local grid or region, smaller loads tend to neutralize to produce a stable time series but for an individual home load, the time series volatility is quite high, thus accurate forecasting becomes challenging task.

## CONCLUSIONS AND FUTURE WORKS

In this paper, we presented an approach to forecast electricity load on individual household level what can potentially provide greater intelligence in smart metering systems. The result of MARS model used for 24 hours ahead short term load forecast shows that it has good performance and reasonable prediction accuracy can be achieved.

Accurate forecasting brings value added both, to a utility provider and individual customers. The first one can plan the resources and also to take control actions to balance the electricity supply and demand. The customers can benefit from metering solutions through greater understanding of their own energy consumption and future projections, allowing them to better manage costs of their usage.

As future work we see the need to undertake home appliance recognition problem under the nonintrusive appliance load monitoring (NIALM) concept. It may generate additional value to smart meters since the electricity usage of a household changes over time based on the operation of various appliances used by the family. Therefore, appliance detection might be additional input variable used for more accurate electricity usage forecasting.

## REFERENCES

Alfares H.K., Nazeeruddin M. (2002) Electric load forecasting: literature survey and classification of method, International Journal of Systems Science, vol. 33(1), 3–34.

Beccali M., Cellura M., Brano V.L., Marvuglia A. (2004) Forecasting daily urban electric load profiles using artificial neural networks, Energy Conversion and Management, vol. 45, 2879–2900.

Brockwell P.J., Davis R.A. (2002) Introduction to Time Series and Forecasting, Springer.

Castillo E., Guijarro B., Alonso M. (2001) Electricity Load Forecast using Functional Networks, Report for EUNITE 2001 Competition, available at http://neuron.tuke.sk/competition/ on 2013-07-10.

Friedman J.H. (1991) Multivariate Adaptive Regression Splines, The Annals of Statistics, vol. 19, 1–141.

Hippert H.S, Pedreira C.E., Souza R.C. (2001) Neural networks for short term load forecasting: a review and evaluation, IEEE Transactions on Power Systems, vol. 16, 44–55.

Javed F., Arshad N., Wallin F., Vassileva I., Dahlquist E. (2012) 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. 69, 150–160.

Khotanzad A., Zhou E., Elragal H. (2002) A neuro-fuzzy approach to short-term load forecasting in a price-sensitive environment, IEEE Transactions on Power Systems, vol. 17, 1273–1282.

Song K.B., Baek Y.S., Hong D.H., Jang G. (2005) Short-term load forecasting for the holidays using fuzzy linear regression method, IEEE Transactions on Power Systems, vol. 20, 96–101.

Weron R. (2006) Modeling and forecasting electricity loads and prices: A statistical approach, Wiley, Chichester.