Analysis of energy market using *data mining* methods

*Beata Gołębiewska, Jędrzej Trajer**

Department of Fundamental Engineering, Warsaw University of Life Sciences
Nowoursynowska Str. 166, 02-787 Warsaw, *corresponding author: jedrzej_trajer@sggw.pl

Received January 9.2013; accepted March 14.2013

Summary. The paper compares selected *Data Mining* techniques for forecasting and describes electricity market including the structure and forms of energy trading. One of the goals was to analyze data for electricity production in Poland. The analysis aimed at studying the influence of selected variables on the examined problem and the effectiveness of the proposed models. Four *Data Mining* methods were used to develop the forecast models, namely, regression and time series using *MLP* neural networks as well as Support Vector Machine (*SVM)* and *MAR-Splines*. Two time horizons of the forecasts, namely, 1 day and 7 days, were studied. The results were used to verify the models and select the best one.

Key words: electricity market, forecast, *Data Mining*.

INTRODUCTION

Electricity market is defined by three key characteristics. One of these characteristics that can be attributed to the specificity of energy, which cannot be stored in large quantities, is the significance of balancing demand and supply, and the fact that electricity overproduction may lead to serious financial losses. Another characteristic is the fact that the energy used by a given consumer cannot be traced back to its source of production. The third characteristic is the variability of demand for energy in time, and its dependence on numerous factors (technology, weather, season of the year, industry).

At present, three types of energy trading can be distinguished on the Polish market:

– *Contract market* is the least complicated in terms of energy trading. It is a place, where contracts between participants of the market are concluded. These contracts specify the price of energy for the period of each day of contract duration. The prices may be fixed, which, in certain situations, may be unprofitable for sellers, or variable, based on demand and instantaneous electricity prices.

– *Exchange market,* where future contracts as well as “day ahead” and “hour ahead” contracts for the purchase and sale of energy are concluded on the Polish Power Exchange (Towarowa Giełda Energii, TGE). Trading is realized on the Day-Ahead Market. Sale and purchase of energy takes place 24 hours before the physical delivery of energy. Orders are placed for the specified hours of the next 24h day. This is the basis for setting hourly electricity prices.

– *Balancing market* is an extremely significant element of the market, whose function is to ensure smooth electricity supply. This function is realized by reacting to demand for energy, and generating the required volume of energy. This sector is managed by the Transmission System Operator – PSE-Operator S.A. – the company which sells or purchases energy to balance supply and demand.

The volume of energy, for which consumers pay, is determined based on forecasts. If the estimated and the actual energy consumption differ, and there is excess of energy, the surplus may be resold on the balancing market. Conversely, in case of deficiency, additional volume of energy may be purchased on the balancing market. Balancing is managed by trading companies and not consumers. Fig. 1 presents a graphical interpretation of this procedure.

The facts presented above clearly show the necessity of developing a model facilitating accurate forecasts of estimated energy consumption, essential for the generation of the required volume of electrical energy, and consequently, more effective balancing of supply and demand [5]. There are various methods to develop forecasting models. New technologies [10], especially neural networks [3, 9] and fuzzy logic [11] are especially useful in this respect. Modifications of currently used forecasting models also contribute to the improvement of forecasts effectiveness [1]. Data Mining methods proved to be an extremely effi-

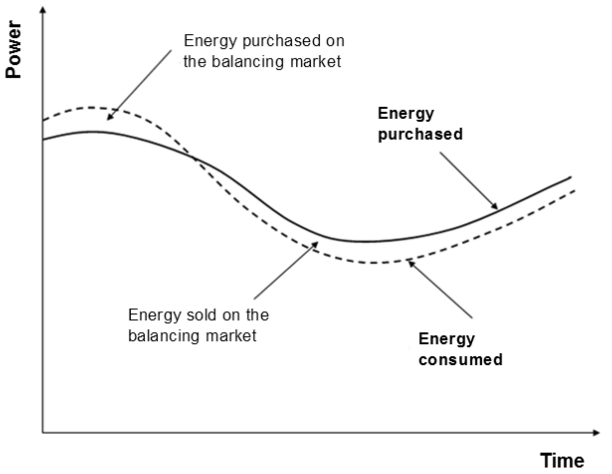
Power
Energy purchased on the balancing market
Energy purchased
Energy sold on the balancing market
Energy consumed
Time

Fig. 1. A graph presenting the dependence of power on time showing how balancing market operates. *Source: www.cire.pl*

cient tool in forecasting, i.a., energy consumption [5]. *Data Mining* methods were used to compare different forecasting models, as they can be applied to on-line analysis of large datasets.

PURPOSE AND METHODS

Data Mining comprises a set of analytical tools and methods that may be applied for various purposes [8]. It relies on computers to discover patterns and regularities in large datasets [2]. These methods can be divided according to a number of criteria. One classification divides data mining methods into supervised learning (learning with a teacher) or unsupervised learning (learning without a teacher) [6]. The first group of methods is used when the situation has been recognized, and the results are included in the training sample. This is the basis for the development of the model, which is then used to classify and categorize new data. The following example might be given to illustrate such a situation: it is necessary to group new consumers according to the criteria specified for each group. Unsupervised methods are used when some data are available but the mechanisms that generate them are unknown. In this case, the goal is to develop the best possible model. These methods, as opposed to supervised methods, are not aimed at discovering relations between variables, but are used to forecast certain values, recognize a situation or present a structure of a phenomenon [7]. An example may be specifying a group of customers according to defined criteria e.g. volume of consumed energy or payment promptness.

The authors analyze the effectiveness of four developed forecasting models that are based on regression and time series using MLP neural networks and support vector machine (*SVM)* as well as *MARSpline*s. The models were developed in the *Statistica 9 Data Miner*. Forecasts for two time horizons were performed for each of these methods, namely, a forecast for the day ahead and day seven. The analysis was performed on the basis of data obtained from *Monthly reports on the operation of the Polish Power System and Balancing Market* (which contain data collected during operative technical planning and during the operation of the system) for the period of five years (2007–2011) The dataset contains 1818 observations (1818 successive days), Fig. 2.

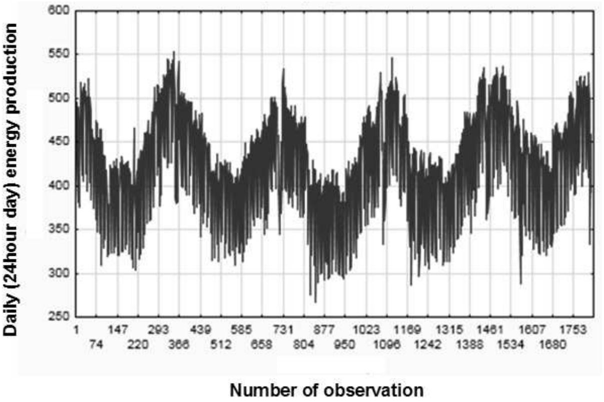

Fig. 2. A plot illustrating variations in daily (24hour day) electricity production [GWh] in the years 2007-2011 *Source: prepared by the authors*

The dataset was divided into two subsets: the training set (80% of cases) and the test set for the verification of the models (20% of cases). The following variables were used to describe these observations:

1. Number of observation (No).
2. Year (Y).
3. Day of year (DY).
4. Number of month (MNo).
5. Number of week (WNo).
6. Day of month (DM).
7. Daily (24hour day) production (DP).
8. Daily production of energy delayed by 1 observation (DP-1).
9. Daily production of energy delayed by 6 observations (DP-6).
10. Daily production of energy delayed by 7 observation (DP-7).
11. Energy production forecast for 1 day ahead (DP+1).
12. Energy production forecast for 7 days ahead (DP+7).
13. A variable informing about holidays (including Sundays) and work days (DS/DR).

The structure of the forecasting models was shown in Fig. 1. Forecasts for all models were prepared for one day and seven days ahead.

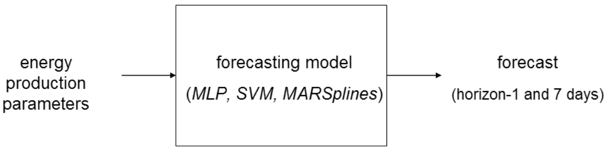

Fig. 3. Structure of forecasting models

Input variables and the structure of each *MLP* model were selected based on the validity analysis (susceptibility analysis) of these models and simulations. In the time series model, only past values for electricity production are considered, and they are selected depending on the horizon and feasible network structure. In contrast, the *SVM* model aims at finding such a basic function that will predict the values of the dependent variable for new observations with the greatest accuracy. The process involves minimization of the error function using various basis functions (radical functions) and appropriate selection of constants *C* and ε in the process of learning. *MARSpline* (*Multivariate Adaptive Regression Splines*), which belongs to Data Mining methods, is a non-parametric method used for solving classification and regression problems. This method does not require preliminary establishment of assumptions concerning the functional relationship between the independent and dependent variables (e.g. linear or logistic relationship). This relationship is constructed from a set of coefficients and basic functions generated from the available data. MARSplines can be used to solve problems with multiple variables at the input, which is likely to cause problems in other methods. The input space is divided into separate regions, each with its own regression function. *MARSplines* algorithm in the Statistica 9 program permits the use of any number of variables, observations and output variables, and consists of two stages. The first stage involves building a model, i.e. increasing its complexity by adding basis functions, and the second stage, called pruning, involves removal of the least significant basis functions. This procedure stops when the *Generalized Cross Validation error* (GCV) has reached its minimal value. The GCV error is a measure of the goodness of fit of the model that takes into account not only the residual error but the model complexity as well.

RESULTS AND DISCUSSION

Input variables for the time series neural model were 21 past values for the daily energy production. Simulations were performed to construct networks with the optimal structure for forecast horizon 1 – *MLP 21:29:1* and 7 – *MLP* 21:43:1. The forecasting model based on MLP networks was developed for the following input variables: number of month and week of year, daily production of energy – current and delayed by 1 and n days (for forecast horizon 1 – n=6 and 7 – n=7), and the qualitative variable that distinguishes work day from holiday). Optimal network structures were constructed for these variables: for forecast horizon 1 – *MLP 6:3:1*, and 7 – *MLP 6:9:1*. Independent variables for forecast horizons 1 and 7 in the *SVM* networks and in *MARSplines* were assumed analogically as in the previous case. In the *SVM* method, optimal parameters were selected using a cross validation sample. Default regression of type 1 was applied, and optimal parameters were obtained with the following values: capacity C= 3 and ε=0.1. Attempts to construct the model proved that the best results were obtained for the radical basis function *RBF*, where the parameter gamma equaled 0.146. Development of *MARSplines* requires that a number of important parameters be set carefully. For instance, a parameter *Maximum number of basis functions* may be found in the specification of the model. This parameter is best set at maximum possible number of basis functions, so that the program can find all the functions. In this case, the number, determined in the course of model construction, was set to 110. Further increase of model complexity may be achieved by increasing the *Order of interactions* between input variables. Number 2 was selected for the analysis involving forecast of energy for the day ahead, which allowed to take into account not only the results but also the interactions between pairs of variables. This parameter was also determined on the basis of preliminary analyses. With this method, pruning is recommended to limit the complexity of the model. It is performed by selecting the option *Remove the least significant basis functions*. This prevents overfitting. Generalized cross validation errors for the model for time horizon of 1 day and 7 days were equal $GCV_1 = 408.7$ and $GCV_7 = 448.9$, respectively. Table 1 presents the results obtained from MLP time series models, and table 2 shows the results for the other examined MLP models.

The developed models were verified using the generated test set. The values of forecast errors and correlation coefficients indicate the model quality. The forecasting models were verified using the following quality assessment measures: correlation coefficient and the values of forecast errors. To facilitate the model quality assessment, it was assumed that the output variable was within the range 267–553 [GWh] for two horizons. The results of the verification of the models for the test sample are presented in Table 3.

The data show that as the length of forecast horizon increases, the quality of the models decreases. Analyzing the results for forecast horizon equal 7 days it can be observed that the highest correlation coefficient is equal to 0.920 and the smallest forecast errors were for neural networks (solv-

Table 1. Forecast results obtained from MLP time series models.

Forecast horizon	Name of network	Quality (learning)	Quality (testing)	Learning error	Test error	Activation (latent)	Activation (output)
1	MLP 21-29-1	0.960	0.956	110.24	138.86	Exponential	Tanh
7	MLP 21-43 -1	0.890	0.900	297.04	302.30	Logistic	Logistic

Table 2. Forecast results obtained from MLP regression models.

Forecast horizon	Name of network	Quality (learning)	Quality (testing)	Learning error	Test error	Activation (latent)	Activation (output)
1	MLP 6-3-1	0.941	0.934	171.51	193.58	Exponential	Tanh
7	MLP 6-9-1	0.937	0.920	177.83	246.35	Exponential	Tanh

Table 3. Quality measures for each model for the test sample.

	Neural networks (Regression)		Neural networks (Time series)		Support Vector Machine		MARSplines	
Forecast horizon [days]	1	7	1	7	1	7	1	7
Average of squared residuals	387.15	492.6903	266.77	609.64	427.59	518.1768	418.27	643.23
Absolute mean error	13.47	14.9002	11.10	17.50	13.99	15.4277	13.88	16.15
Correlation coefficient	0.932	0.920	0.956	0.900	0.926	0.906	0.926	0.884

ing regression problems). The worst results were obtained using the *MARSplines* method, where the correlation coefficient was equal to 0.884 and the values of forecast errors were the highest.

CONCLUSIONS

Research to develop the most effective forecasting methods, in particular for short forecast horizons, has been conducted by many research centres. *Data Mining* methods are very useful for that purpose, and allow to prepare forecasts of electricity production in a short time. However, the best results (in the two examined horizons) were obtained for regression using MPL neural networks, with the smallest difference between them. Moreover, the results confirm that the developed MLP models yield accurate predictions for new data (small forecast error for the test set).

The advantage of the last studied method, MARSplines, is the fact that it does not require preliminary assumptions. It does well for outlier observations and, consequently, better characterizes the examined phenomenon. It also gives better results in case of higher number of explanatory variables. This method also has its own quality measure (goodness of fit), taking into account the complexity of the model and the residual error. Nevertheless, despite numerous advantages, the interpretation of results in this method is likely to create difficulties. Like in other non-parametric methods, the implementation of the model in this method may cause problems.

REFERENCES

1. **Assimakopoulos V, Nikolopoulos K., 2000:** *The theta model: a decomposition approach to forecasting.* Elsevier – International Journal of Forecasting 16 (2000) 521–530.
2. **Berry M., Linoff G., 2000:** *Mastering Data Mining.* Wiley, Hoboken, New York.
3. **Crone S., Hibon M., Nikolopoulos K., 2011:** *Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction..* International Journal of Forecasting 27, 635–660.
4. **Nęcka K., 2011a:** *Analysis of the Continuity of Electric Energy Supply in Poland.* TEKA Kom. Mot. i Energ. Roln. – OL PAN, 11c, 230–236.
5. **Nęcka K., 2011b:** *Use of Data Mining Techniques for Predicting Electric Energy Demand.* TEKA Kom. Mot. i Energ. Roln. – OL PAN, 2011, 11c, 237–245.
6. **Migut G., 2009:** *Czy stosowanie metod Data Mining może przynieść korzyści w badaniach naukowych*, StatSoft Polska.
7. **Sokołowski A., Pasztyła A., 2004:** *Data mining w prognozowaniu zapotrzebowania na nośniki energii*, Akademia Ekonomiczna w Krakowie, StatSoft Polska.
8. **Tadeusiewicz R., 2006:** *Data mining jako szansa na relatywnie tanie dokonywanie odkryć naukowych poprzez przekopywanie pozornie całkowicie wyeksploatowanych danych empirycznych.* Dostępny w internecie: http://www.statsoft.pl/
9. **Trajer J., Czekalski D., 2011:** *Prognozowanie sum promieniowania słonecznego.* Polska Energetyka Słoneczna,2–4, 39–41.
10. **Trojanowska M., 2008:** *.Alternative Methods of Estimating Rural Consumers' Daily Demand for Electrical Energy.* TEKA Kom. Mot. Energ. Roln. – OL PAN, 8, 287–291.
11. **Trojanowska M.,Małopolski J., 2011:** *Forecast Models of Electric Energy Consumption by Village Recipients over a Long-Term Horizon Based on Fuzzy Logic.* TEKA Kom. Mot. i Energ. Roln. – OL PAN, 2011, 11c, 327–334.
12. Źródło internetowe: *www.cire.pl*

ANALIZA RYNKU ENERGETYCZNEGO
Z WYKORZYSTANIEM METOD 'DATA MINING'

Streszczenie. W pracy przedstawiono techniki *Data Mining* służące do prognozowania oraz istotę rynku energii elektrycznej uwzględniająca strukturę i formy handlu energią. Celem pracy była analiza danych dotyczących produkcji energii elektrycznej w Polsce. Analiza miała na celu zbadanie istoty wybranych zmiennych oraz na badane zjawisko oraz skuteczności zaproponowanych modeli. Modele prognostyczne zostały zbudowane za pomocą czterech metod *Data Mining*: regresja i szeregi czasowe z wykorzystaniem sieci neuronowych typu *MLP* oraz metoda wektorów nośnych *SVM* i *MARSplines*. W każdej z metod zostały uwzględnione dwa horyzonty czasowe prognoz 1 i 7 dni. Otrzymane wyniki posłużyły do weryfikacji modeli i wyboru najlepszego.

Słowa kluczowe. rynek energii elektrycznej, prognozowanie, *Data Mining*.