**Krzysztof NĘCKA**
Agricultural University of Cracow, Department of Power Engineering and Agricultural Processes Automation
Balicka Str. 116B, 30-149 Kraków, Poland
e-mail: krzysztof.necka@ur.krakow.pl

# THE USE OF MACHINE LEARNING TECHNIQUE FOR SHORT-TERM FORECASTING OF DEMAND FOR ELECTRICITY

*Summary*

*The study verifies the usefulness of selected machine learning techniques for predicting hourly demand for electricity within a short time period. The results of the performed analyses show that the lowest values for both the MAPE forecast error for the test set at the level of 17% and the lowest share of the balancing energy in the total consumption at a level which does not exceed 15% were obtained for models for which the input data included the averaged electricity consumption profile for characteristic days of the week, the forecast number of pure production pieces and the encoded day of the week and time of the day. Among the tested models, forecasts prepared on the basis of artificial neural networks and standard CRT trees were characterised by the best quality of predictions.*
***Key words:*** *data mining, electricity, machine learning, short-term forecast.*

# WYKORZYSTANIE TECHNIK UCZENIA MASZYNOWEGO DO KRÓTKOTERMINOWEGO PROGNOZOWANIA ZAPOTRZEBOWANIA NA ENERGIĘ ELEKTRYCZNĄ

*Streszczenie*

*W pracy sprawdzono przydatność wybranych technik uczenia maszynowego do predykcji godzinowego zapotrzebowania na energię elektryczną w krótkim horyzoncie czasu. Z wykonanych analiz wynika, że najniższe wartości zarówno błędu prognozy MAPE na poziomie 17% jak i najniższy udział energii bilansującej w całkowitym zużyciu na poziomie nie przekraczającym 15% uzyskano dla modeli, dla których zmiennymi wejściowymi były uśredniony profil zużycia energii elektrycznej dla charakterystycznych dni tygodnia, prognozowana liczba sztuk czystej produkcji oraz zakodowany dzień tygodnia i godzina doby. Spośród badanych modeli najlepszą jakością predykcji charakteryzowały się prognozy opracowywane w oparciu o sztuczne sieci neuronowe oraz standardowe drzewa CRT.*
***Słowa kluczowe:*** *data mining, energia elektryczna, prognoza krótkoterminowa, uczenie maszynowe.*

## 1. Introduction

Electricity has been one of the best known and desired forms of energy for many years. The guarantee of its availability in the required quantity and with appropriate quality is necessary for proper functioning and development of each society [1, 2]. Even short interruptions in its supply or failure to keep the requirements regarding its quality results in enormous financial losses for companies which have to restrict or even stop the production process [3-5]. They also result in lowering residents' standard of living as the majority of household devices require electricity to work.

Since 1 July 2007, together with complete opening of the electricity market, for each recipients, regardless on the annual electricity consumption [6], forecasting the demand for electric power over a short period of time has become a very important issue. For scheduled recipients, correct determination of the planned consumption of electric power at individual times of day influences the cost of electricity, which constitutes a large share in total costs of its operation [7]. However, the determination of a contract item with the lowest possible error at individual times of a commercial day is a very challenging issue, also due to the fact that it is affected by a lot factors whose influence is not obvious.

Currently, due to the decreasing costs of devices allowing for continuous electricity consumption registrations, access to basic data enabling forecasting has become very easy, even for small recipients. However, a problem appears how from a large database containing, apart from information about the electric power consumption, also information about the technological process and, e.g. weather conditions, to get the data allowing for creating an accurate forecast. Proper identification of exogenous variables is possible owing to the use of available methods [8-11]. Among many known methods, machine learning techniques seem to be very effective, especially for large sets of data. Their name is derived from the main characteristic of these algorithms i.e. multiple, interactive searching of the data set to find an effective model or hidden patterns. The assessment of the quality of the model built is not made on the basis of statistical significance but only on the basis of the prediction correctness calculated for the test set. Such an assessment method simulates actual operating conditions of the model when new exogenous data are entered. An additional advantage of these techniques includes the fact that no formal assumptions must be met concerning the form of the function, distribution of variables or constant variance.

## 2. The aim, object and methodology of the research

The aim of the study was a comparative analysis of the application of selected machine learning methods used for forecasting the hourly demand for electric power within a short time period at a food and agriculture production plant.

The research was performed at a company with its registered office in the Małopolska province. This was a fam-

ily-run company operating on the market since 1990. It has a modern poultry slaughter line together with a cold store. Its basic activities include slaughter and sales of poultry on the domestic market as well as in Europe, Asia and Africa.
For the purposes of the forecasting system construction, developed with maximum adjustment of the model describing the process of the demand for electricity to the specificity of each recipient, the suitability of the following models was checked: classical regression ones, regression trees, models using artificial neural networks (ANN) as well as multivariate adaptive regression using splines (MARSplines).

The hourly demand for electricity was estimated using the aforementioned methods, taking into account previously determined statistical relationships between the demand for energy and the electricity consumption in analogous previous periods, weather conditions and the production size [12].

The problem of the quality of the developed forecast is one of the more important issues concerning the forecasts. Within the quality assessment of the developed forecasts, their admissibility and accuracy was verified by determining indices frequently used in the literature:

- the mean forecast error:

$$ME = \frac{1}{n} \cdot \sum_{t=1}^{n} E_t - E_t^* , \quad (1)$$

- mean relative forecast error:

$$MAPE = \frac{1}{n} \cdot \sum_{t=1}^{n} \frac{\left| E_t - E_t^* \right|}{E_t} \cdot 100\% , \quad (2)$$

- the share of actual amount of the balancing energy in the total energy consumption:

$$\%E = \frac{\sum_{t=1}^{n} \left| E_t - E_t^* \right|}{\sum_{t=1}^{n} E_t} \cdot 100\% , \quad (3)$$

where:
$E_t$ - the actual quantity of electricity used at time $t$,
$E_t^*$ - the forecast quantity of electricity used at time $t$,
$n$ - the number of last observation of the forecast variable.

## 3. Research results

Before the commencement of the construction of forecasting models, the input data were divided into two sets. The first one, which was a learning set, was created from 6937 observations recorded in the first research period, i.e. from December 2011 to October 2012. The other set, the so-called test set, was created from 1680 observations recorded during the last period of the research. The time horizon for each forecasts prepared was 48 hours due to the requirements imposed on electricity market participants in Poland.

For the development of prediction models allowing for determination of hourly demand for electricity, a project was created in the graphic environment of *Statistica 10 Data Miner*, whose working area is divided into four sections:

– *Data sources*, in which the file location containing data intended for application in the further modelling process.
– *Data preparation, cleaning and transformation.* In this section, the data are separated into the learning and test sets on the basis of the encoding variable recorded in the prepared input file.
– *Data analysis*: The following models were used in the constructed project template allowing for forecasting the size of the hourly demand for electricity: general stepwise regression, standard regression trees *CRT*, exhaustive *CHAID* for regression, enhanced regression trees, artificial neural networks (*ANN*) Multivariate Adaptive Regression Splines (*MARS*).
– *Reports*: Results summing up individual analyses in the form of file folders are placed in this part of the working area for *Data Mining* projects after the completion of calculations.

While testing the models, the following sets of exogenous variables were examined:

A. electricity consumption smoothed out using the 4253H filter for the entire plant with a 168-hour delay and the forecast mean daily temperature outside;
B. the planned (on the basis of constructs signed with poultry suppliers) number of pure production pieces, the forecast mean daily temperature outside and the encoded day of the week and time of the day (33 variables in total, including 31 “0”-“1” variables);
C. the product of the forecast (on the basis of constructs signed with poultry suppliers) number of pure production pieces and the encoded time of the day and the encoded day of the week (31 variables in total, including 7 “0”-“1” variables);
D. the forecast (on the basis of constructs signed with poultry suppliers) number of pure production pieces, the forecast daily temperature outside and the encoded day of the week and time of the day (32 variables in total, including 31 “0”-“1” variables). At this point no direct hourly demand for energy was forecast as in the remaining cases. During the first step, the mean hourly load profile for individual days of the week was determined. Next, using the actual flow, deviations from the mean profile were calculated during individual hours. On the basis of the aforementioned explanatory variables, a model of deviations from the mean profile was constructed and a forecast of hourly demand for energy was determined for the test set as a sum of the profile load and the forecast deviation.

After launching the constructed algorithm on the basis of the selected methods, the program automatically performs the requested numerical analyses and determines forecasts of hourly demand for electricity. During the construction of all forecasting models, no changes were made to the division of cases into the learning and test sets created at the beginning.

## 4. Assessment of forecasting models

Within the quality assessment of the prepared forecasts, their accuracy was verified by determining characteristic indices calculated for the test set (Table 1).

All models constructed on the basis of the first set of explanatory variables in the form of electricity consumption smoothed out using the 4253H filter for the entire plant with a 168-hour delay and the forecast mean daily temperature outside were characterised by a high MAPE error oscillating around 30%. Also, high values ranging from 25.9%

(the CHAID model) to 28.2% (the CRT model) were obtained for the index characterising the share of the actual quantity of balancing energy in the total energy consumption.

Table 1. Comparison of the quality of forecasts generated by means of selected methods on the basis of the set explanatory variables "A"
*Tab. 1. Porównanie jakości prognoz generowanych wybranymi metodami na podstawie zestawu zmiennych objaśniających „A"*

| The method: | Error measure | | |
|---|---|---|---|
| | ME [kWh] | MAPE [%] | %E [%] |
| general stepwise regression | -6,0 | 32,7 | 28,0 |
| standard CRT trees | -5,3 | 32,8 | 28,2 |
| exhaustive CHAID | -3,3 | 29,3 | 25,9 |
| enhanced regression trees | -6,7 | 32,9 | 27,8 |
| SSN | -1 | 29,2 | 26,1 |
| MARS | -4,8 | 32,4 | 28,0 |

*Source: Own study / Źródło: opracowanie własne*

The introduction of explanatory variables in the form of the forecast number of pure production pieces, the temperature outside and the day of the week and the time of the day in the encoded form resulted in a reduction in the MAPE forecast error to 20-25%. Also, a decrease by approx. 10% in the share of the balancing energy (ΔESR) was observed in the total energy consumption for the entire test period. The CRT and ANN models were characterised by the lowest forecasting errors (Table 2).

Table 2. Comparison of the quality of forecasts generated using selected methods on the basis of the set explanatory variables "B"
*Tab. 2. Porównanie jakości prognoz generowanych wybranymi metodami na podstawie zestawu zmiennych objaśniających „B"*

| The method: | Error measure | | |
|---|---|---|---|
| | ME [kWh] | MAPE [%] | %E [%] |
| general stepwise regression | -3,0 | 23,7 | 20,5 |
| standard CRT trees | -4,4 | 20,4 | 16,9 |
| exhaustive CHAID | -7,8 | 24,0 | 19,6 |
| enhanced regression trees | -3,7 | 25,1 | 21,9 |
| SSN | -5,7 | 19,3 | 16,1 |
| MARS | -3,4 | 24,6 | 21,7 |

*Source: Own study / Źródło: opracowanie własne*

The construction of models based on aggregated input variables in the form of the product of the forecast number of pure production pieces and the encoded time of the day as well as seven variables depicting the encoded day of the week allowed for a further reduction in the errors of hourly electricity consumption for the facility under analysis calculated for the test set. For the constructed models, the MAPE error value ranged from 18.2% for the ANN model to 22.9% for the CHAID model, which corresponded to a 15.3% and 19% share of the balancing energy, respectively (Table 3).

Table 3. Comparison of the quality of forecasts generated by means of selected methods on the basis of the set explanatory variables "C"
*Tab. 3. Porównanie jakości prognoz generowanych wybranymi metodami na podstawie zestawu zmiennych objaśniających „C"*

| The method: | Error measure | | |
|---|---|---|---|
| | ME [kWh] | MAPE [%] | %E [%] |
| general stepwise regression | -4,0 | 20,2 | 16,6 |
| standard CRT trees | -3,6 | 20,4 | 16,9 |
| exhaustive CHAID | -6,8 | 22,9 | 19,0 |
| enhanced regression trees | -5,0 | 22,4 | 18,9 |
| SSN | -3,0 | 18,2 | 15,3 |
| MARS | -4,2 | 22,6 | 18,8 |

*Source: Own study / Źródło: opracowanie własne*

The lowest values of indices characterising the quality of the constructed forecasting models were obtained for the following variables: the electricity consumption profile for characteristic days of the week, the forecast number of pure production pieces and the encoded day of the week and time of the day The value of the forecasting error calculated on the test set was approx. 16-17% and it was nearly twice as low as the analogous index determined for models constructed on the basis of electricity consumption for the entire plant with a 168-hour delay and the forecast mean daily temperature outside (Table 4).

Table 4. Comparison of the quality of forecasts generated by means of selected methods on the basis of the set explanatory variables "D"
*Tab. 4. Porównanie jakości prognoz generowanych wybranymi metodami na podstawie zestawu zmiennych objaśniających „D"*

| The method: | Error measure | | |
|---|---|---|---|
| | ME [kWh] | MAPE [%] | %E [%] |
| general stepwise regression | -3,0 | 17,4 | 15,0 |
| standard CRT trees | -2,7 | 16,8 | 14,5 |
| exhaustive CHAID | -5,3 | 17,2 | 14,7 |
| enhanced regression trees | -2,6 | 16,6 | 14,4 |
| SSN | -3,1 | 16,4 | 14,1 |
| MARS | -3,4 | 17,4 | 15,0 |

*Source: Own study / Źródło: opracowanie własne*

## 5. Conclusions

1. The results of the analyses performed show that the lowest values for the MAPE forecast error at the level of approx. 16-17% were obtained for models for which the input data included the averaged electricity consumption profile for characteristic days of the week, the forecast number of pure production pieces and the encoded day of the week and time of the day. For these predictors, also the lowest quantities of balancing energy were obtained whose share in the total consumption did not exceed 15%.
2. Among the tested classical and alternative models, forecasts built on the basis of artificial neural networks were characterised by the best quality of predictions for the test

set. Standard CRT trees were ranked second in terms of the quality of forecasts. However, this method proved to be very sensitive to the set of input data as for 4 various sets of variables it rendered the best forecasts twice and the worst forecast once. Among the tested models, forecasts encumbered by the greatest errors were obtained for the MARS method.

3. A very significant advantage of forecasting models constructed in the *Data Miner* graphic environment is the fact that in order to perform analyses again, e.g. after changing the parameters of individual models or after adding new data, it is enough to launch the previously constructed project template and the entire computational process will be performed automatically.

## 6. References


[1] Nęcka K., Trojanowska M.: Programming of rural power networks development Part II. Draft guidelines for the local plan of energy supply. TEKA Komisji Motoryzacji i Energetyki Rolnictwa, Vol. X. Lublin, 2010, s. 294-300.

[2] Polityka energetyczna Polski do 2030 roku. Załącznik do uchwały nr 202/2009 Rady Ministrów z dnia 10 listopada 2009 r.

[3] Motowidlak T.: Istota ciągłości dostaw energii elektrycznej w Unii Europejskiej. Polityka Energetyczna, 2007, Tom 10, Zeszyt 1. PL ISSN 1429-6675.

[4] Hanzelka, Z.: Koszty dostawy złej jakości energii elektrycznej. Automatyka, Elektryka, Zakłócenia, 2012, nr 7, s. 11-19..

[5] Popczyk J.: Nowe spojrzenie na jakość energii elektrycznej w Polsce w warunkach urynkowienia elektroenergetyki i integracji z UE, Przegląd Elektrotechniczny, 2004, R. 80, nr 6, s. 568-571.

[6] Ustawa z dnia 10 kwietnia 1997 r. Prawo energetyczne. Dz.U. z 1997 r. nr 54 poz. 348 wraz z późniejszymi zmianami.

[7] Ciepiela D.: Koszty bilansowania – zmora klienta. [online]. [dostęp 27–08–2013]. Dostępny w Internecie: http: //energetyka.wnp.pl/tpa/poradnik_jak_zmienic dostawce energii/koszty-bilansowania-zmora-klienta,3359_2_0_1.html.

[8] Trojanowska M.: Analiza zapotrzebowania na moc i energię elektryczną w zakładzie mleczarskim. Journal of Research and Applications in Agricultural Engineering, 2010, 55(2), 113-116.

[9] Nęcka K.: Analiza sezonowości obciążeń w zakładzie przemysłu rolno-spożywczego. Technika Rolnicza Ogrodnicza Leśna. 2011, 3, 25-26.

[10] Piotrowski P.: Prognozowanie krótkoterminowe godzinowych obciążeń w spółce dystrybucyjnej z wykorzystaniem sieci neuronowych – analiza wpływu doboru i przetworzenia danych na jakość prognoz. Przegląd Elektrotechniczny, 2007, 83, nr 7-8, 40-43.

[11] Weron R., Misiorek A.: Zwiększenie dokładności prognoz ceny energii poprzez zastosowanie preprocessingu oraz modeli nieliniowych. Przegląd Elektrotechniczny, LXXXII, 2006, nr 9, 44-46.

[12] Nęcka K.: Wpływ wstępnego przetwarzania danych wejściowych na jakość modeli predykcyjnych budowanych technikami Data Mining. Materiały konferencyjne. Prognozowanie w Elektroenergetyce PE, 2013.