BADANIA OPERACYJNE I DECYZJE

Nr 4 2008

Barbara GŁADYSZ*, Dorota KUCHTA*

APPLICATION OF REGRESSION TREES IN THE ANALYSIS OF ELECTRICITY LOAD

In the paper electricity load analysis was performed for a power region in Poland. Identifying the factors that influence the electricity demand and determining the nature of the influence is a crucial element of an effective energy management. In order to analyse the electricity load level the CART (Classification and Regression Tree) method has been used. The data for the analysis are hourly observations of the electricity load and weather throughout one year period. Two categories of factors were taken as predictor variables, on which the demand for the electricity load depends: variables describing weather and variables representing structure days in a year. An analysis of the errors of the presented models was carried out.

Keywords: *data mining, classification and regression tree, electricity load, weather*

1. Introduction

The distributors of electric energy have to manage it in such a way that the costs of electricity buying and selling are minimized. The costs depend on the decision taken about the current electricity demand. Therefore, identifying the factors that influence the electricity demand and determining the nature of the influence is a crucial element of effective energy management.

If we analyse the nature of electricity demand it is clear that it is characterized by cyclicality, seasonality and randomness. In power engineering three types of seasonality can be distinguished: the yearly, the weekly and the daily-night cycles. The weekly cycle is the result of a week's work cycle. The year cycle is essentially influenced by climate conditions and scholar-holiday time. The daily-night cycle is the result of both climate conditions and daily work time. The number and complexity of

* Institute of Organization and Management, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland, e-mail: Barbara.Gladysz@pwr.wroc.pl, Dorota.Kuchta@pwr.wroc.pl

the factors which influence this special stochastic model result in numerous short-term forecasting models and techniques used, such as autoregression models, exponential smoothing models, seasonal decomposition models, multifactor econometric models, neutral networks, fuzzy methods. A review of the methods used in the analysis of electricity demand can be found in [1], [10]. Especially important are the techniques which take into account weather factors. We will mention here only several selected forecasting models which take into account meteorological factors, because the literature in this domain is very vast.

What can be applied here is classical regression with independent variables such that the system load in previous periods, the type of calendar day, the temperature [7]. Ružić et al. proposes the regression with quality variable taking on three levels, which describe temperature ranges important from the point of view of the energy demand [14]. Wójciak and Wójcicka use seasonal model (SARIMA) for forecasting energy demand, by constructing forecasts in three stages [15]. In the first stage, they eliminate the seasonality linked to the calendar, by determining the averages for one name periods. In the following stage, they take into account meteorological factors (temperature, cloudiness, sunset time) while modeling the differences between the standard energy demand curve and the actual energy demand. In the third stage, the residuals gained from the average daily values under the consideration of the atmospheric conditions are analyzed by means of ARIMA. Lotufo and Minusi build up forecasting models of the energy load separately for the summer and the winter periods [10]. The forecasts are constructed in two stages. In the first one, ARIMA models are built on the whole observations set, in the second one – for the models residuals ARIMA constructs a regression which takes into account the temperature. Another approach is the robust regression. Gładysz and Kuchta use robust regression with the temperature as the independent variable [5], [6].

Also fuzzy models can be used to forecast the energy demand. Gładysz builds up the fuzzy regression taking into account the energy system load in previous periods, the calendar day type and the temperature [8]. In [2] the authors put forward a fuzzy model – in the model the load is a time varying function and takes the form of Fourier's coefficients. The weather input is limited only to temperature deviation. Chenthur Pandian et al. use fuzzy logic to model energy demand [4]. They assume that both the forecast (the energy load) and the independent variables (the day time and the temperature) are fuzzy numbers of the second type.

Mastorocostas uses a constrained orthogonal least-squares method for generating TSK fuzzy models [12]. As forecast factors he assumes the energy demand in previous periods and the temperature.

Also neuron networks find a vast use in forecasting the energy system load. While constructing the forecasts, the following factors are assumed e.g., the average daily load of the given day from previous years, the season (winter, summer, spring, autumn) and the day type (a holiday, a working day) [13].

In the paper, in order to analyse the electricity load level the regression tree method has been used. Our aim is to investigate the influence of atmospheric conditions on the electricity system load at various moments during 24 hours. The data for the analysis are hourly observations of the electricity load and weather throughout one year.

2. Analysis of change in the electricity load

The CART (Classification and Regression Tree) method is a data exploration method which can be used for big data sets. It was proposed by Breiman et al. [3]. At present there exist many variants of this method [9], [11]. The goal of the regression tree technique is determining the factors influencing the explored feature – decision variable – and the nature of this influence. The dependency between the factors and the explored feature is described in the form of decision rules expressed in the tree structure. The tree is a kind of graph constituting a set of branches going from a node called root to end nodes, called leaves. The branches of the tree are decision nodes. Each branch goes from the root or a decision node to another node, being another decision node or a leaf, thus the final decision. Decision trees built by means of the CART technique are strictly binary. Each decision node is a beginning of two subbranches. To select the variables in the tree the variance minimization criterion was used.

The regression trees have been constructed for the electricity load at 1 o'clock a.m., 7 o'clock a.m., 12 o'clock noon and at 7 o'clock p.m. for a power region in Poland. This choice of hours is a consequence of the electricity level structure throughout the 24 hour period: using the *k*-means method [9] to classify the electricity load through the 24 hour period the following clusters were obtained: night hours, morning hours, working hours, afternoon hours and evening hours (see Table 1).

Table 1. Clasters

C	a.m.											noon	p.m.											midnight
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
Two	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1
Three	1	1	1	1	1	1	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3
Four	1	1	1	1	1	1	3	2	2	2	2	2	2	2	2	2	2	2	4	4	2	2	2	2

To select the clasters, L^2 norm was used:

$$\sum_{c=1}^{C}\sum_{j=1}^{n_c}(y_{jc}-\overline{y}_c)^2$$

where:

y_{jc} – the value of decision variable Y for the j-th observation in the c-th claster,

$\overline{y}_c$ – the mean point of the decision variable Y of the observation in the c-th claster.

For the construction of regression trees by means of the CART method for the distinguished hours two categories of factors were taken as possible predictor variables, on which depends the demand for the electric energy. In the category daily weather data we distinguished:

TMIN – minimal temperature [°F],
TMEAN – mean temperature [°F],
TMAX – maximal temperature [°F],
DPMIN – minimal due point [°F],
DPMEAN – mean due point [°F],
DPMAX – maximal due point [°F],
HMIN – minimal humidity [%],
HMEAN – mean humidity [%],
HMAX – maximal humidity [%],
PMIN – minimal sea level pressure [in Hg],
PMEAN – mean sea level pressure [in Hg],
WMEAN – mean wind speed [km/h],
WMAX – maximal wind speed [km/h],
F – fall (A – any, F – fog, FR – fog-rain, FRS – fog-rain-snow, FS – fog-snow, FRT – fog-rain-thunderstorm, R – rain, RS – rain-snow, RST – rain-snow-thunderstorm, RT – rain-thunderstorm, S – snow).

In the category type of day in the year we assumed:

TIMEW winter time (1 – winter time, 0 – the other days),
TIMES summer time (1– summer time, 0 – the other days),
DWEEK – day of week (1 – Monday, 2 – Tuesday, 3 – Wednesday, 4 – Thursday, 5 – Friday, 6 – Saturday, 7 – Sunday),
GROUP – week days groups (1 – Saturday, 2 – Sunday, 3 – Monday, 4 – Tuesday, Wednesday, Thursday, Friday),
SUNDAY – Sunday (1 – Sunday, 0 – the other days),
HOL – holiday (1 – holiday, 0 – the other days),
HOLA – a day after a holiday (1 – a day after a holiday, 0 – the other days),
HOLB – a day before a holiday (1 – a day before a holiday, 0 – the other days),
DF – a free day(1 – a free day (bank holiday), 0 – the other days),
DB – a day between holidays (1 – a day between holidays, 0 – the other days),
SHOLW – winter school holiday (1 – winter school holiday, 0 – the other days),
SHOLS – summer school holiday (1 – summer school holiday, 0 – the other days),
SHOL – a school holiday (1 – a school holiday, 0 – the other days).

In subsequent iterations we choose as the factor determining the value of decision variable *Y* the one for which the following function attains its minimal value:

$$MSE = \sum_{l=1}^{L_t} \frac{1}{n_{lt}} \sum_{i=1}^{n_{lt}} (y_{ilt} - \bar{y}_{lt})^2$$

where:

y_{ilt} – the value of decision variable *Y* for the *i*-th observation in the *l*-th child of node *t*,

$\bar{y}_{lt}$ – estimator of the decision variables value = the average value of the decision variable *Y* for the observation in the *l*-th child of node *t*,

n_{lt} – observations number in the *l*-th child of node *t*,

L_t – number of children into which node *t* is divided,

$n_t = \sum_{l=1}^{L_t} n_{lt}$ – observations number in node *t*.

The regression trees presented in the paper have been determined by means of the SAS package. The analysis results are presented in figures 1, 2, 3, 4.

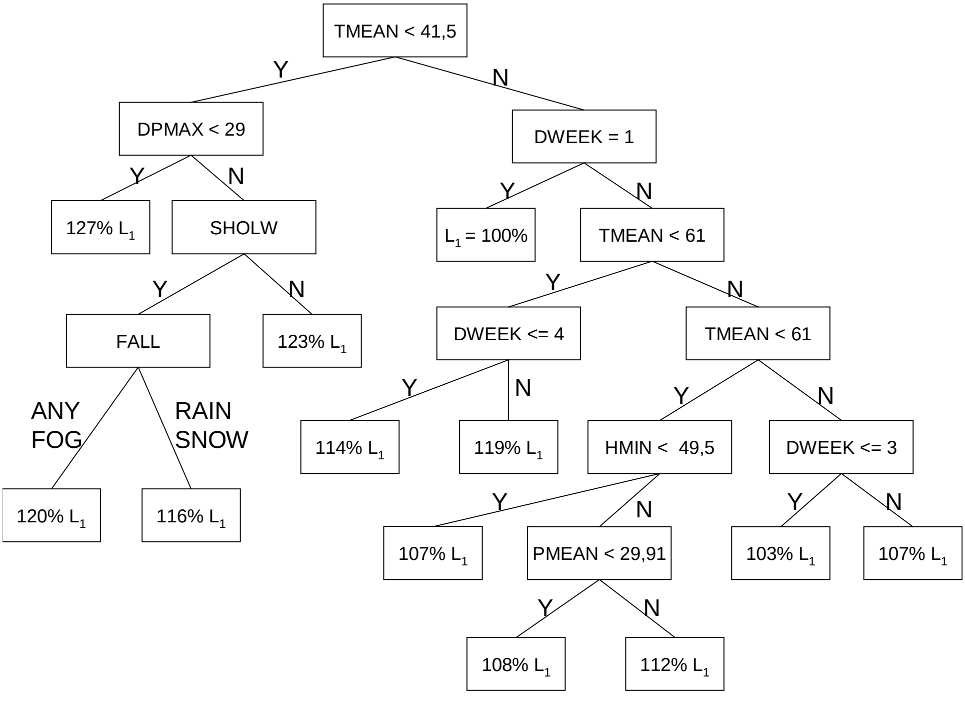

Fig. 1. Regression tree for electricity load at 1 o’clock a.m.

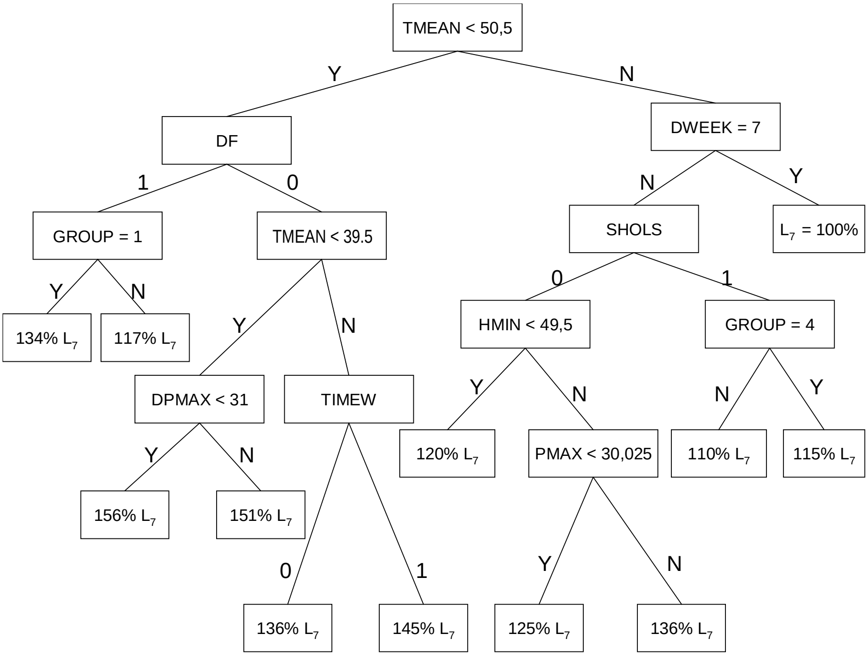

Fig. 2. Regression tree for electricity load at 7 o'clock a.m.

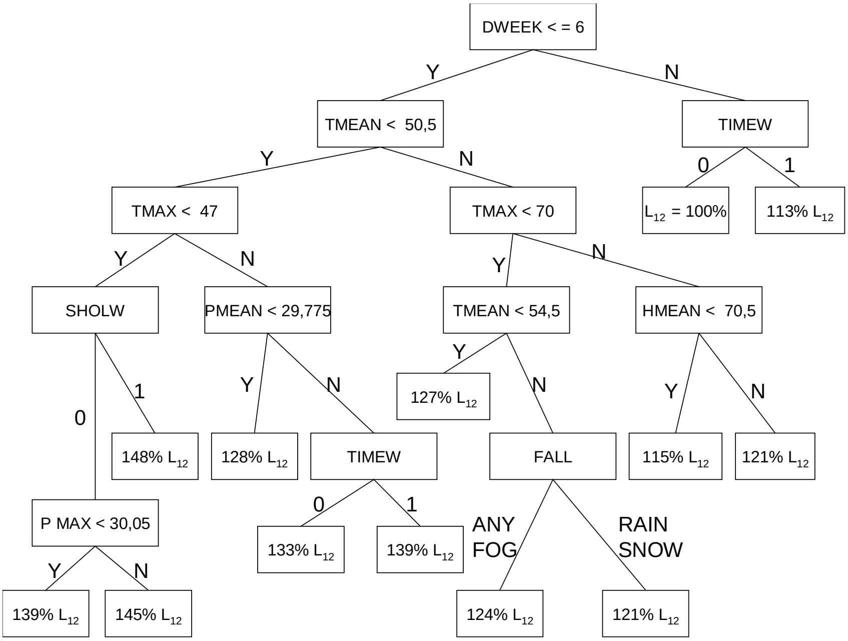

Fig. 3. Regression tree for electricity load at 12 o'clock noon

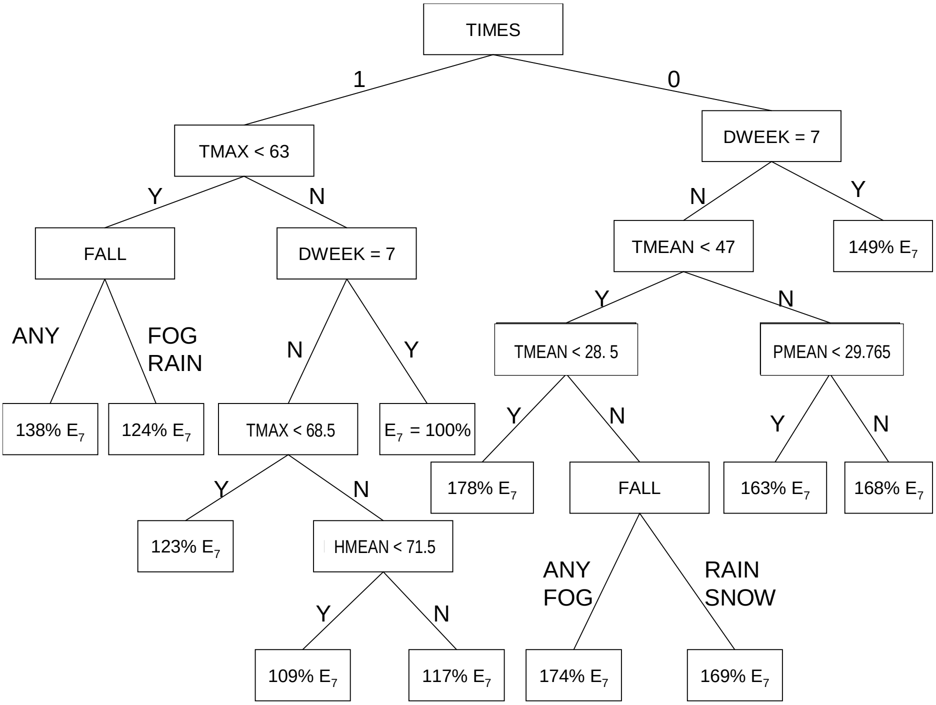

Fig. 4. Regression tree for electricity load at 7 o'clock p.m.

In Table 2 the factors having a substantial influence on the electricity load in the analysed moments during the 24 hours period and the threshold values of those factors are given. Also the meteorological falls (or their lack) influence the electricity load level (1 o'clock a.m., 12 o'clock noon, 7 o'clock p.m.). Table 2 also includes measures of the regression trees quality: the variability coefficient for the whole tree $(\sqrt{MSE} / \bar{y})$ and the determination coefficient R^2. The lowest value of the determination coefficient (0.67) has the regression tree for 1 o'clock a.m. for the other hours the value of R^2 is between 0.74–0.78. The CART analysis distinguished as the substantial factors influencing the electricity load: temperature, dew point, humidity, pressure and type of fall. The wind speed turned out to be a meteorological factor without an essential influence on the electricity load level. Among the factors of the category type of day the following ones turned out to be essential: day of the week, week days group, summer and winter school holidays, and also summer and winter times. In the night and morning hours the mean temperature is essential, in the evening hours the maximal temperature, and during the afternoon the energetic load depends both on the mean and the maximal temperature. Dew point influences the level of electricity load in the night and morning hours. Additionally, the electricity load depends on the minimal humidity (7 o'clock a.m.) and mean humidity (the other hours) and on the mean sea level pressure (1 o’clock a.m., 12 o’clock noon, 7 o’clock p.m.) as well as on the maximal sea level pressure (7 o’clock a.m., 12 o’clock noon).

Table 2. Factors influencing the electricity load distinguished by means of the CART method and the measures of the regression trees quality

Hour	T_MEAN	T_MAX	DP_MAX	H_MIN	H_MEAN	P_MEAN	P_MAX	F	Type of day	Coefficient of variation	R^2
	[ºF]	[ºF]	[ºF]	[%]	[%]	[in Hg]	[in Hg]			[%]	
1 o’clock a.m.	41.5 50.5 61		29		49.5	29.91		yes	D_WEEK SHOL_W	4.8	0,67
7 o’clock a.m.	39.5 50.5		31	49.5			30.025	no	D_WEEK GROUP DF SHOL_S TIME_W	6.8	0,76
12 at noon	50.5 54.5	47 70			70.5	29.775	30.05	yes	D_WEEK SHOL_W TIME_W	5.3	0,74
7 o’clock p.m.		47 63 68.5			71.5	29.765		no	D_WEEK TIME_S	5.7	0.78

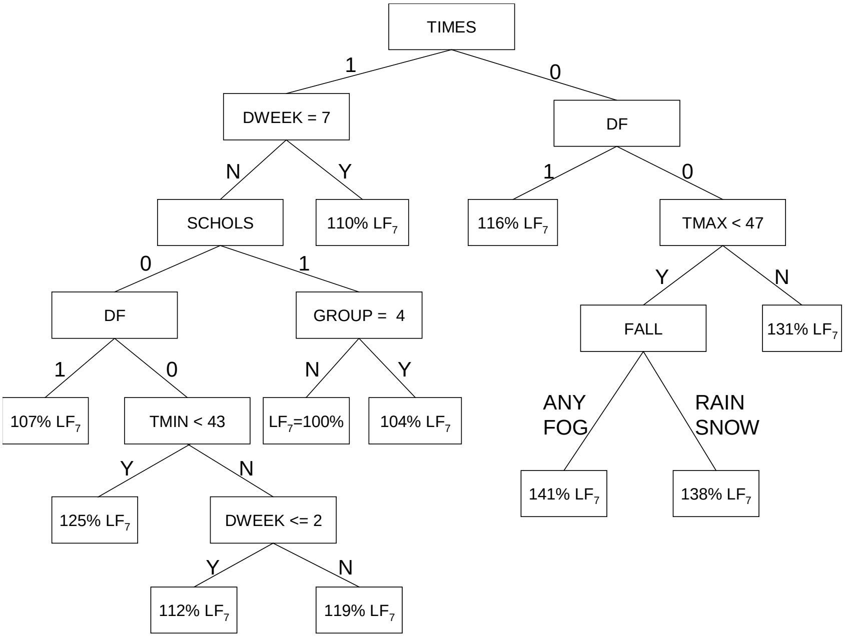

Fig. 5. Regression tree for electricity load at 7 o’clock a.m. with weather factors: temperature and kind of fall

We also constructed regression trees taking into account only two types of meteorological factors: temperature and kind of fall. The resulting trees had similar variability parameters and determination coefficients. An example of such a tree – for 7 o'clock a.m. – is presented in figure 5.

3. Conclusions

The CART method showed that – along with the work cycle and the scholar cycle – it is the atmospheric conditions which influence substantially the electricity load level. The classification and regression tree analysis distinguished as the substantial factors influencing the electricity load temperature, dew point, humidity, pressure and type of fall. The wind speed turned out to be a meteorological factor without an essential influence on the electricity load. Among the factors of the category "type of day" the following ones turned out to be essential: day of week, week days group, summer and winter school holidays, and also summer and winter times.

As far as sources for constructing the forecasts are concerned, we can use the meteorological forecast data worked out in meteorological institutions. There is a relatively easy access to these weather data: they are published on web pages, e.g. www.pogoda.onet.pl, www.wunderground.com. We can also forecast the values of individual meteorological variables by means of advanced techniques used in meteorology.

It should be emphasized that the errors of the CART models were estimated for known weather data values. If the weather data are biased with an error, it will influence the error of energy load forecast.

References

[1] ALFARES H.K., NAZEERUDDIN M., *Electric load forecasting: literature survey and classification of methods*, International Journal of Systems Science, 2002, 33, 23–34.

[2] AL-KANDARI A.M., SOLIMAN S.A., EL-HAWARY M.E., *Fuzzy short-term load forecasting*, Electrical Power and Energy Systems, 2004, 26.

[3] BREIMAN L., FRIEDMAN J., OLSEN R., STONE CH, *Classification and Regression Trees*, Chapman and Hall/CRC Press, Boca Raton, Fl, 1984.

[4] CHENTHUR PANDIAN S., DURAISWAMY K., CHRISTOBER ASIR RAJAN C., KANAGARAJ N., *Fuzzy approach for short term load forecasting*, Electric Power and Systems Research, 2006, 76.

[5] GŁADYSZ B., *Regresja odporna w prognozowaniu obciążenia elektroenergetycznego* [in:] P. Dittmann, J. Szanduła (eds.), *Prognozowanie w zarządzaniu firmą*, Indygo Zahir Media, Wrocław 2008, 120–126.

[6] GŁADYSZ B., KUCHTA D., *Outliers detection in selected fuzzy regression models* [in:] *Application of Fuzzy Sets Theory*, F. Massulli, S. Mitra, G. Pasi (eds.), 7-th International workshop on Fuzzy Logic, WILF 2007, Camogli Italy, July 2007, Proceedings, LNAI 2007, Springer-Verlag, Berlin–Heidelberg 2007.

[7] GŁADYSZ B., KOŁWZAN W., MERCIK J., *Prognozowanie w zarządzaniu energią* [in:] *Prognozowanie w zarządzaniu firmą*, P. Dittmann, J. Krupowicz (eds.), Wydawnictwo Akademii Ekonomicznej, Wrocław 2006, 197–206.

[8] GŁADYSZ B., *The electric power load fuzzy regression model* [in:] *Issues in Soft Computing Decisions of Operations Research*, O. Hryniewicz, J. Kacprzyk, D. Kuchta (eds.), Academic Publishing House EXIT, Warszawa 2005, 171–180.

[9] LAROSE D.T., *Discovering Knowledge in Data*, An Introduction to Data Mining, John Wiley and Sons, Inc., Indianapolis 2005.

[10] LOTUFO A.D.P., MINUSSI C.R., *Electric power systems load forecasting. A survey*, Proceedings of IEEE Power Tech'99 Conference, Budapest 1999.

[11] KANTARDZIC M., *Data mining: Concepts, Models, Methods, and Algorithms*, Wiley-Interscience, Hoboken, New York 2003.

[12] MASTOROCOSTAS P.A., THEOCHARIS J.B., PETRIDIS S.P., *A constrained orthogonal least-squares method for generating TSK fuzzy models: Application to short-term load forecasting*, Fuzzy Sets and Systems, 2001, 118, 215–233.

[13] OSOWSKI S., *Sieci neuronowe do przetwarzania informacji*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2000.

[14] RUŽIĆ S., VUČKOVIĆ A., NIKOLIĆ N., *Weather sensitive method for short term load forecasting in electric power utility of Serbia*, IEEE Transactions on Power Systems, 2003, 18, 1581–1586.

[15] WÓJCIAK M., WÓJCICKA A., *O metodzie korekty prognoz minimalizującej koszt niezbilansowania zapotrzebowania na energię elektryczną* [in:] *Prognozowanie w zarządzaniu firmą*, P. Dittmann, J. Krupowicz (eds.), Wydawnictwo Akademii Ekonomicznej, Wrocław 2006, 336–347.

Drzewa regresyjne w analizie obciążenia systemu elektroenergetycznego

W artykule zbadano wpływ warunków atmosferycznych na poziom obciążenia systemu elektroenergetycznego. Identyfikacja czynników warunkujących wielkość popytu na energię elektryczną jest podstawowym elementem systemu zarządzania energią elektryczną. W badaniach zastosowano metodę *k*-średnich oraz technikę drzew klasyfikacyjnych i regresyjnych. W pierwszym etapie badań metodą *k*-średnich wyróżniono jednorodne – z uwagi na obciążenie systemu elektroenergetycznego – grupy godzin w skali doby. Dla każdej z grup dla wybranej godziny (reprezentanta) zbudowano drzewo regresyjne, przyjmując jako czynniki dane meteorologiczne oraz typ dnia w skali roku. Przeprowadzona analiza pokazała, że czynnikami warunkującymi poziom obciążenia systemu energetycznego są: temperatura, punkt rosy, wilgotność oraz rodzaj opadów. Informacja o rodzaju oraz wartościach progowych tych czynników meteorologicznych może zostać wykorzystana w procesie prognozowania poziomu obciążenia systemu elektroenergetycznego i tym samym przyczynić się do poprawy efektywności procesów zarządzania energią. Przeprowadzono analizę błędów skonstruowanych drzew regresyjnych.

Słowa kluczowe: *analiza danych, metoda k-średnich, drzewa klasyfikacyjne i regresyjne, obciążenie systemu elektroenergetycznego, pogoda*