Optimizing the Acquisition Cost of Input Data for Daily National Power System Load Forecasts Using Automated Statistical Methods

Author
Rafał Czapaj

Keywords
NPS load, NPS power demand, hourly average forecasts, explanatory variables, input parameters, meteorological parameters, statistical methods, data mining

Abstract
The paper presents the possibility of using statistical methods to automate the selection of explanatory variables to balance the daily load of the National Power System (NPS). With automation, the cost of input forecast purchase may be optimized by minimizing their number, and the results also allow for a reduction in the effort required to select input parameters (explanatory variables) for later forecasting of NPS daily loads.

DOI: 10.12736/issn.2300-3022.2017303

Received: 13.02.2017
Accepted: 08.03.2017
Available online: 30.09.2017

1. Introduction
One of the power system security determinants is the accuracy of power system demand forecasting [1]. The transmission grid operator bears many risks, including that of the significant deviation of forecasts from the National Power System's actual load [2]. The best choice of explanatory variables (input parameters) and of an effective statistical method is the crucial step of building a predictive model. By improving the match of explanatory variables to dependent variables, the precision of the description of the dependent variables is enhanced [3]. Careful selection of explanatory variables is the key to building the best predictive model. However, this selection is often highly labour-intensive (time-intensive), and obtaining precise, complete and reliable historical data can be difficult and involves considerable financial effort [4]. The ideal situation, from the point of view of the preparation of the forecast for the power system operator, is the minimization of workload and (very often) of the difficulties and costs of the acquisition of data constituting a set of explanatory variables. The labour-intensity, handling data in the form of explanatory variables, and their acquisition cost can be minimized with:

- automated processes of the identification and ranking of the variables most favourable to a process (using statistical software)
- selection of the best explanatory variables from a given acquired data group – by way of their testing by various statistical methods.

The acquisition cost of explanatory variables which are meteorological measurements (historical data and forecasts) used for business and industrial purposes is one of the highest costs. Meteorological parameters specific to Poland's geographical location [5, 6] significantly influence the NPS load.

Depending on the analysed time interval (time series), the optimum forecasting method selection is also necessary in the process of predictive model development, in addition to the selection of explanatory variables [7]. Their choice greatly impacts the results and largely depends on the expertise of the forecaster. In the predictive model development process, explanatory variables and/or forecasting methods are selected again when, in particular, a previous model fails to perform satisfactorily [8].

The simulations reported in this publication attempted to use an automated method for selecting the best explanatory variables. The best explanatory variables are those that describe the explained variable as accurately as possible. The best explanatory variables were identified in connection with the historical data of 15-minute peak power demand in a day in the NPS system using selected statistical methods. In the next step, the potential cost was calculated of their annual purchase for the hourly resolution

of daily load. This allowed for the assessment of the capacity for limiting the number of parameters of the weekly forecasts with hourly data points which may be bought weekly. The ideal situation is where with one input parameter (explanatory variable), a forecast of the analysed parameter can be developed. But more than one parameter affects the NPS load, so it seems that any reduction in capital expenditure and effort for the final forecast preparation may be beneficial for the forecaster.

The presented approach to optimizing the number of predictive input parameters of the model may be useful for three user groups:

- experienced researcher-forecasters whose predictive models (methods) have become immune to the changes observed in nature (they respond with a delay to the dynamics of their behaviour)
- beginner researcher-forecasters, who have a basic knowledge of data preparation for forecasts and forecasting, who operate with limited time resources
- acquisition cost managers of input data for forecasting.

The NPS' 15-minute maximum (peak) power demand in a day was selected as the explained (forecast) variable [9]. The predictive model was fed with the historical details of individual parameters and then, after specifying the optimal number of input parameters, the annual cost was simulated of purchasing the forecasts of selected parameters. It was assumed that the best solution to the problem was to minimise the effort and capital expenditures for input data acquisition for forecasting.

Meteorological explanatory variables	Unit	Code
Max. ambient temperature	°C	Zm6
Min. ambient temperature	°C	Zm7
Rainfall	mm	Zm8
Average wind speed	km/h	Zm9
Average wind speed (based on expert statistical intervention	km/h	Zm10
Max. wind speed	km/h	Zm11
Atmospheric pressure	hPa	Zm12
Number of heating degree-days	°C x day	Zm13
Number of cooling degree-days	°C x day	Zm14
Number of sunshine hours	–	Zm15
Solar energy	W/m^2	Zm16
UV radiation	–	Zm17
Dew point temperature	°C	Zm18
Wet thermometer temperature	°C	Zm19

Tab. 1. Examples of meteorological parameters that are explanatory variables and may be purchased

2. Meteorological explanatory variables (externalvariables)

For the purpose of this study, meteorological measurements from a location in the south of Poland were considered a good approximation of the average meteorological conditions in the

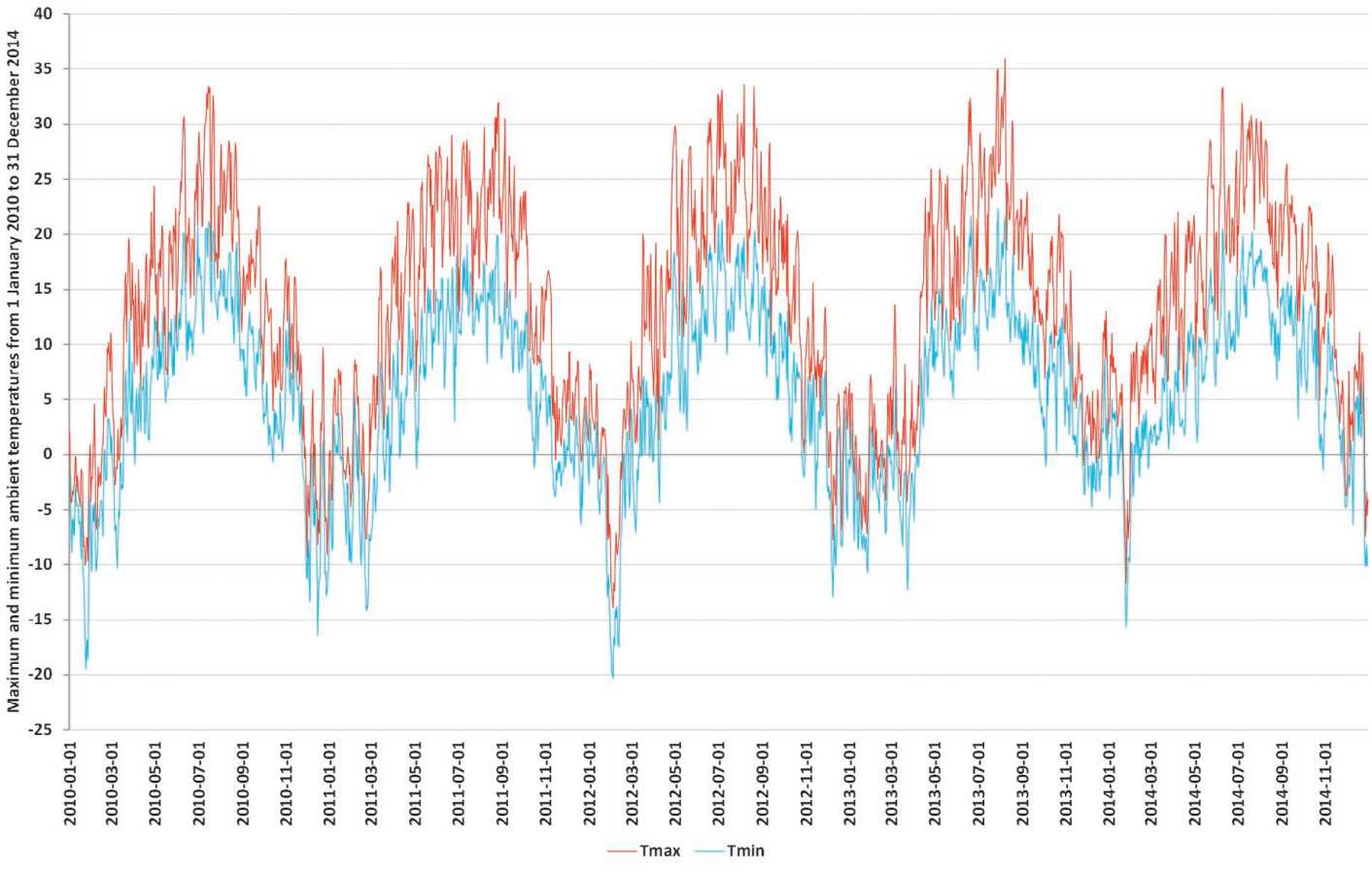


Fig. 1. Maximum and minimum daily ambient temperatures in five calendar years

entire NPS and selected as the set of meteorological explanatory variables. This location is neither the coldest location (Suwałki) nor the warmest location (Wroclaw/Legnica), a comparison of observations from this station with historical data from several other weather databases reveal that it is a rough approximation of the arithmetic mean (ca. ±1°C) from both temperature poles. The first set of variables (the most numerous) consisted of the following 14 meteorological parameters, which may be purchased from specialized vendors (Tab. 1).

In Fig. 1 the maximum and minimum daily ambient temperatures at this location are shown in the period from the beginning of January 2010 to the end of December 2014.

Two explanatory variables are added to the above set, to complement it. The first of them (Zm2) contains in its coded form the details of the date of measurement (year/month/day) with a distinction made between consecutive days of the week and the split between non-holiday and holiday days. The other variable (Zm3) is a non-coded time form, which contains details of time (with a 15-minute resolution) of the 15-minute peak power NPS load, and wind generation output to the NPS [9] in each day of the analysed time series.

3. Explanatory variables (external variables)

The variable explained in the section concerning the identification of explanatory variables was the 15-minute peak power of the daily NPS load. Later simulations were carried out for forecasting the average hourly load of the NPS in the whole calendar year (52 weeks).

In addition, in order to highlight the impact of wind conditions on the NPS load, variables Zm20-23 (Tab. 2) were tested, and the suitability of the encoded moon cycle (Zm5) details to explain the dependent variables was assessed. The internal explanatory variables are listed in Tab. 2.

Each explanatory variable is a stream of historical data collected once a day and covering the five-year period from January 1, 2010 to December 31, 2014.

The automated methods for explanatory variable identification included also (subject to the relevant criteria) the methods shown in Tab. 3. Other explanatory variable identification methods included the methods listed in Tab. 4.

With the approach described above, the following sets of explanatory variables were obtained by each method – Tab. 5.

4. Potential costs of the acquisition of input forecasts for the daily NPS load forecast for a year, in hourly intervals

Pre-selected explanatory variable sets are listed in Tab. 6, and their graphical interpretation is shown in Fig. 2.

Analysis of the data from Fig. 2 indicates that out of all the paid variables (red), 3 out of 5 automated methods selected the smallest input variable sets needed to produce the most accurate NPS daily peak load forecasts. From the above three, the least number of input parameters was required by the MARS method and the Pearson method (5 variables). The multiple regression method selected 6 input variables.

Other explanatory variables	Unit	Code
Daily 15-minute peak power share in the weekly peak	%	Zm4
Coded moon phase details	–	Zm5
Maximum wind farm output	MW	Zm20
Hour of maximum wind farm output	–	Zm21
Available wind farm capacity	MW	Zm22
Wind farm output to installed capacity ratio	%	Zm23

Tab. 2. Other parameters which are explanatory variables [9, 10]

Automated methods for explanatory variables identification	Criterion	Code
Multiple regression (classical method)	B > ±0,04	M1
MARS (data mining)	predictor ranking	M2
Pearson coefficient calculation (classical method)	>0.47	M4
Fast C&RT (data mining)	predictor ranking	M8
Selection and elimination of variables (data mining)	variable rank	M10

Tab. 3. Selected automated methods for the identification of explanatory variables

Other explanatory variable identification methods	Criterion	Code
Selection of explanatory variables perceived to have a significant impact on the NPS load	Proprietary	M3
Selection of all explanatory variables	None	M5
Selection of all of the best explanatory variables from M1-M4 methods	According to their respective criteria	M6
Selection of all of the best explanatory variables and expert selection of additional explanatory variables	None + proprietary	M7
Multiple regression, iteratively separate for each explained variable	B > ±0,1	M9

Tab. 4. Other explanatory variable identification methods

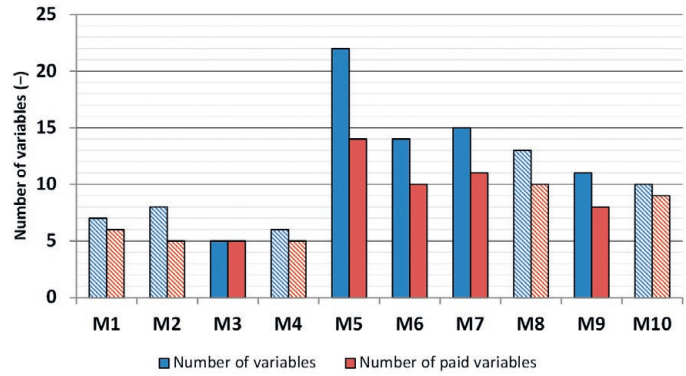
Method	Variables	N.umber of variables	N.umber of paid variables
M1	4, 6–7, 13, 16–17, 19	7	6
M2	3–4, 6–7, 16–17, 20, 22	8	5
M3	11, 13–15, 18	5	5
M4	4, 6–7, 13, 16–17	6	5
M5	1, 3–23	22	14
M6	3–4, 6–7, 11, 13–20, 22	14	10
M7	3–4, 6–7, 11–20, 22 (additional var.: pressure Z12)	15	11
M8	4, 6–7, 12–20, 22	13	10
M9	4, 6–7, 13–14, 16–19, 22–23	11	8
M10	3-4, 6-7, 13-14, 16-19	10	9

Tab. 5. List of explanatory variable sets, all variables and paid variables

In order to simulate the annual forecast purchase cost, four variants (V1-V4) of the acquisition costs of a one week-ahead hourly forecast of a single parameter were assumed to be 100–1,000 PLN.

For this simulation, it was assumed that the forecast will be of the NPS load in every hour of the day in a year, without determining whether it is the maximum, average or minimum. Data used for the cost simulation are listed in Tab. 7 and Fig. 3. In addition, listed in Tab. 7 are arithmetic mean values of the four forecasting effectiveness measures [11]. The measurements are the MPE, MAPE, RMSPE, and Theila coefficients. The percentages of forecast errors in the previous five years were averaged, using the 15 forecasting methods listed in Tab. 8.

Analysis of the data in Tab. 7 and Fig. 3 indicate that the M2, M3 and M4 methods, by selecting the lowest number of pay variables from the total number of explanatory variables, allow for the lowest average acquisition cost of these variables forecasts in a year (52 weeks). Moreover, the M2 method, in addition to minimizing the input variable forecast purchase cost, produced the most *ex post* accurate forecasts in 2010–2014. In addition, the M2 method minimizes the time spent on the development of the forecasting model by allowing the researcher's insight to automate the relevance evaluation of input variables from a set of variables. Thus, the M2 method seems to be the most effective, the cheapest and the fastest in the simulated experiment.



Note! The results of the automated methods are hatched

Fig. 2. Number of explanatory variables in each dataset chosen using the selected variable selection methods

5. Summary

The basic purpose of the study, selected results of which are reported in this paper, was to automate and select the optimal set of explanatory variables with the aim of minimising the capital and labour expenditures. Considering the above, it is proposed to recommend the M1 (multiple regression), M2 (MARS), and M4 (Pearson's coefficient) variable set selection methods. It is also noted that for the proposed set of payables, the smallest volume of data was also selected on the basis of literature research, experience and the author's knowledge. The results obtained by M8 (C&RT) and M10 (variable selection and elimination) *data mining* methods almost doubled the results of the best methods, so more caution should be given to selecting these two methods in the future. The presented approach demonstrates its usefulness in the initial optimization of projected forecasting costs. With the four different pricing scenarios assumed for individual explanatory variables forecasts it may be observed that, regardless of the distribution of the cost of purchase of individual explanatory variable forecasts, M2 (MARS), M3 (Pearson's c.f.) and M4 (manual selection) will produce the lowest yearly-average

Method	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
N.umber of variables	**7**	**8**	5	**6**	22	14	15	**13**	11	**10**
N.umber of paid variables	**6**	**5**	5	**5**	14	10	11	**10**	8	9

Note: Automated methods are bolded

Tab. 6. Numbers of explanatory variables preliminarily selected from a set of 22 input (explanatory) variables

Method	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
N.umber of paid variables (-)	6	5	5	5	14	10	11	10	8	9
Number of weeks (-)	52	52	52	52	52	52	52	52	52	52
Weekly cost of input variable forecast (x1000 PLN) V1	1	1	1	1	1	1	1	1	1	1
Weekly cost of input variable forecast (x1000 PLN) V2	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Weekly cost of input variable forecast (x1000 PLN) V3	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Weekly cost of input variable forecast (x1000 PLN) V4	0.6	0.7	0.8	0.9	1	1	0.9	0.8	0.7	0.6
Annual cost (x1000 PLN) V1	312	260	260	260	728	520	572	520	416	468
Annual cost (x1000 PLN) V2	312	234	208	182	437	260	229	156	83	47
Annual cost (x1000 PLN) V3	31	52	78	104	364	312	400	416	374	468
Annual cost (x1000 PLN) V4	187	182	208	234	728	520	515	416	291	281
W1-W4 annual average cost	**211**	**182**	**189**	**195**	**564**	**403**	**429**	**377**	**291**	**316**
Averaged forecast accuracy estimates (%)	**3.64**	**2.94**	**7.11**	**3.64**	**4.62**	**4.24**	**4.23**	**4.50**	**4.54**	**4.40**

Note: Automated methods are bolded

Tab. 7. Simulation of the annual acquisition cost of input variables forecasts for the short-term forecast of the NPS daily loads in the context of averaged ex-post forecasting accuracy measurements

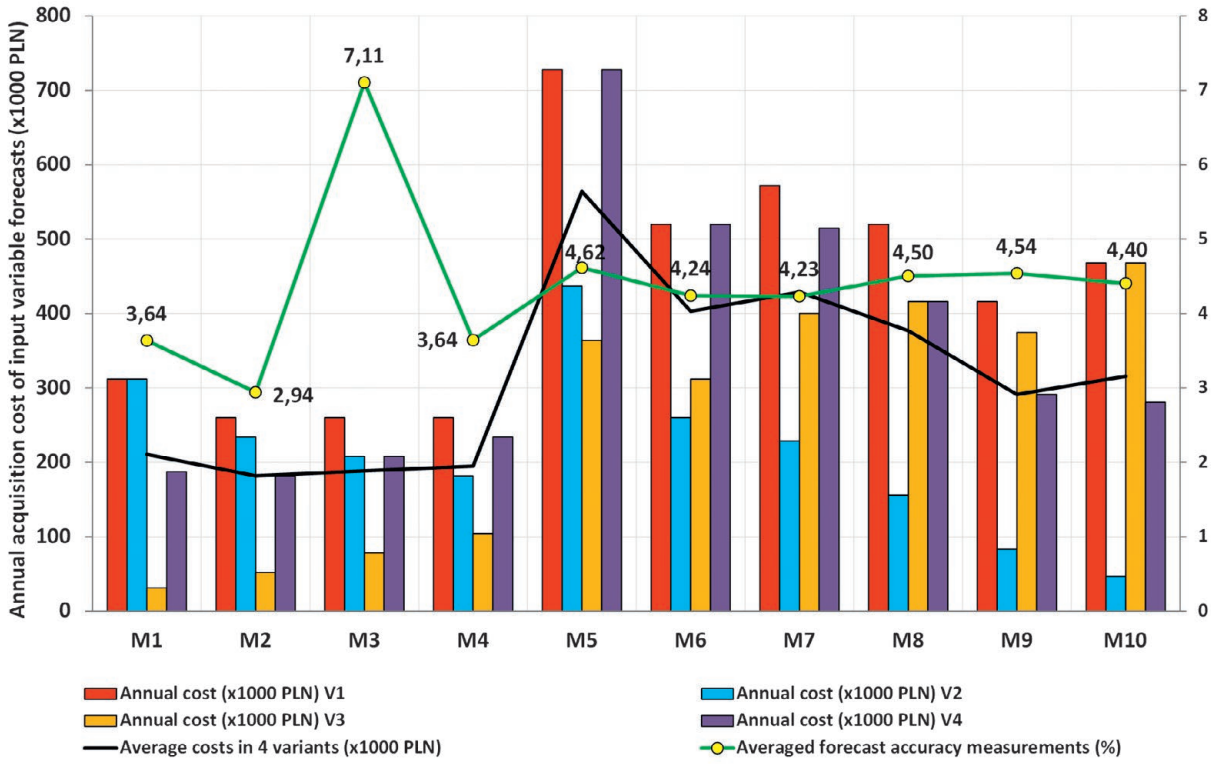


Fig. 3. Simulation of the annual acquisition cost of input variables forecasts for the short-term forecast of the NPS daily loads in the context of averaged ex-post forecasting accuracy measurements

Method	Abbreviation	Code
Multivariate Adaptive Regression Splines	MARS	S1
General models of classification and regression trees (standard version)	C&RT std.	S2
General models of classification and regression trees (version with systems)	C&RT w. sys.	S3
Chi-square automatic interaction detector (standard version)	CHAID std.	S4
Chi-square automatic interaction detector (version with systems)	C&RT w. sys.	S5
Interactive regression trees (interactive version)	C&RT inter.	S6
Interactive regression trees (interactive version)	CHAID inter.	S7
Interactive regression trees (exhaustive version)	CHAID exhaust.	S8
Generalized additive models using the identity binding function	UMA iden.	S9
Generalized additive models using the logarithmic binding function	UMA log.	S10
Multiple regression	Regr. Wlrk.	S11
General linear and nonlinear models	OMLN	S12
General regression models	OMR	S13
(Partial) least squares models	MNKC	S14
Artificial neural networks	SSN	S15

Tab. 8. Statistical methods used for simulation tests

costs of purchase (<200,000 PLN) for the assumed purchase cost variants. The above is due to the fact that each method requested 5 paid variables. This factor has demonstrated that these three methods have a resilience to potential changes in input data purchase costs. The choice of the best out of the three least expensive methods, which in the future could be the most advantageous for building effective forecasts, may be achieved by imposing on Fig. 3 the arithmetic mean of the ex-post performance of forecasts. Out of the three pre-selected methods, M3 must be dropped, because it displayed the worst performance (7.11%) in the five-year period. Out of the remaining two, M2 (MARS) features an outstanding *ex-post* performance, the only one that allowed for a predictive efficiency of less than 3% (2.94%). Therefore, this method minimizes the cost of input variable forecast purchase by providing the researcher-forecaster with the ability to automatically select explanatory variables, and ensures the highest ex-post forecasting efficiency of all the methods reported in this paper. Thus, the MARS method may be considered to be the one that minimizes the time it takes to prepare input data for the forecasting model, minimizes the cost of forecasting the data, and that provides (to the extent of the real conditions of the adopted approach) the best accuracy of forecast results.

It is necessary to verify the performance of the developed sets of explanatory (input) data (variables) for *ex-ante* forecasting of the short-term daily NPS load in hourly intervals. Such verification should include a set of classical and data mining methods.

Please take note of an interesting interdependence that the completed tests and comparative analysis of their results (based on Tab. 6 and 7) indicate that the focus will most likely be on the MARS method, which selected only 5 paid input data (variables) from the group of 8 [11]. This method belongs to the data mining group and in addition to statistical analysis it offers a quick and automated path to obtain the best set of explanatory variables. It is also worth noting that the results are the most favourable in

the 5-year historical data set in question. However, at this stage, it is important to point out that the MARS method at the learning stage (*ex-post* prediction) is susceptible to over-learning, and confronted with real predictions it may produce less promising results.

A valuable experience would be the inclusion, as explanatory variables, of the historic data on electricity prices on the Balancing Market, POLPX, or stock exchange futures, such as intraday and intraday Brent oil prices [12].

REFERENCES

1. G. Bartodziej, M. Tomaszewski, "Polityka energetyczna i bezpieczeństwo energetyczne" [Energy policy and energy security] – issue II, „Nowa Energia" Publishers, Racibórz 2008.
2. A. Weron, R. Weron, "Giełda Energii: strategie zarządzania ryzykiem" [Energy Exchange: risk management strategies], CIRE, Hugo Steinhaus Center of Stochastic Methods, Wrocław University of Technology, Wrocław 2000.
3. R. Klóska, M. Hundert, R. Czyżycki, "Wybrane zagadnienia z prognozowania" [Selected forecasting issues], Economicus, Szczecin 2007.
4. A. Zeliaś, B. Pawełek, S. Wanat, "Prognozowanie ekonomiczne. Teoria, przykłady, zadania" [Economic forecasting. Theory, examples, tasks], Wydawnictwo Naukowe PWN, Warsaw 2003.
5. "Prognozowanie w elektroenergetyce. Zagadnienia wybrane" [Forecasting in power engineering. Selected issues] , editor I. Dobrzyńska, Silesian University of Technology Publishers, Częstochowa 2012.
6. K. Kopecki et al., "Analiza i prognoza obciążeń elektroenergetycznych" [Analysis and forecast of electrical power loads], Wydawnictwa Naukowo-Techniczne, Warsaw 1971.
7. D. Witkowska, "Podstawy ekonometrii i teorii prognozowania. Podręcznik z przykładami i zadaniami" [Fundamentals of econometrics and forecasting theory. A textbook with examples and tasks] , Oficyna Ekonomiczna, Kraków 2005.
8. S.M. Kot, J. Jakubowski, A. Sokołowski, "Statystyka. Podręcznik dla studiów ekonomicznych" [Statistics.Textbook for Economic Students], Centrum Doradztwa i Informacji, Warsaw 2007.
9. www.pse.pl.
10. R. Pieczarko, M. Sołtysik, "Analiza wpływu generacji źródeł wiatrowych na poziom kształtowania się cen energii elektrycznej na rynku SPOT" [Analysis of the impact of wind generation on the electricity pricing in the SPOT market], Scientific Conference „Forecasting in power engineering", Podlesice 2016.
11. R. Czapaj, P. Rzepka, M. Szablicki, "Typowanie zmiennych objaśniających przy wykorzystaniu zautomatyzowanych metod statystycznych jako sposób optymalizacji wyboru metody estymacji szczytowego dobowego obciążenia KSE" [Selection of explanatory variables using automated statistical methods as a way of optimizing the choice of the method of estimating the NPS peak daily load], Scientific Conference „Forecasting in power engineering", Podlesice 2016.
12. M. Kozakiewicz et al., "Zastosowanie ekonometrycznych modeli prognostycznych w transakcjach proprietary trading" [Use of econometric forecasting models in proprietary trading], Scientific Conference "Electricity Market", Kazimierz Dolny 2015.

Rafał Czapaj
PSE Innowacje sp. z o.o.
e-mail: rafal.czapaj@pse.pl
A graduate of the Electrical Engineering Department of Silesian University of Technology (2003). PSE SA Capital Group since January 2005. From 2005 to 2011 in EPC SA dealing with issues of the electricity market, and technical and economic analyses. Since 2011 in PSE Innowacje sp. z o.o. (formerly CATA), dealing with the same subject matter.

This is a supporting translation of the original text published in this issue of "Acta Energetica" on pages 37–42. When referring to the article please refer to the original text.

PL

Optymalizacja kosztów zakupu danych wejściowych do prognoz dobowego obciążenia KSE przy wykorzystaniu zautomatyzowanych metod statystycznych

Autor
Rafał Czapaj

Słowa kluczowe
obciążenie KSE, zapotrzebowanie mocy KSE, prognozy średnich wartości godzinowych, zmienne objaśniające, parametry wejściowe, parametry meteorologiczne, metody statystyczne, *data mining*

Streszczenie
Artykuł prezentuje możliwość skorzystania z metod statystycznych automatyzujących dobór zmiennych objaśniających na przykładzie dobowego obciążenia Krajowego Systemu Elektroenergetycznego (KSE). Automatyzacja pozwala na optymalizację kosztów zakupu prognoz wejściowych dzięki minimalizacji ich liczby, a uzyskane wyniki pozwalają dodatkowo na zmniejszenie nakładów pracy związanych z wyborem parametrów wejściowych (zmiennych objaśniających) na potrzeby późniejszego opracowywania prognoz dobowego obciążenia KSE.

Data wpływu do redakcji: 13.02.2017
Data akceptacji artykułu: 08.03.2017
Data publikacji online: 30.09.2017

1. Wstęp
Jedną z determinant bezpieczeństwa systemu elektroenergetycznego (SEE) jest dokładność prognozowania zapotrzebowania SEE na moc [1]. Operator sieci przesyłowej ponosi wiele ryzyk, do których należy m.in. ryzyko znaczącego odchylenia prognozy od rzeczywistego obciążenia Krajowego Systemu Elektroenergetycznego (KSE) [2]. Jak najlepszy dobór zmiennych objaśniających (parametrów wejściowych) oraz skutecznej metody statystycznej stanowi kluczowy etap w procesie budowy modelu prognostycznego. Ze wzrostem jakości dopasowania zmiennych objaśniających do zmiennej objaśnianej rośnie precyzja opisu zmiennej objaśnianej przez te zmienne [3]. Staranny dobór zmiennych objaśniających jest kluczem do zbudowania jak najlepszego modelu prognostycznego. Dobór ten wymaga jednak często dużych nakładów pracy (czasochłonność), a pozyskiwanie pełnych i wiarygodnych danych historycznych, charakteryzujących się wysoką rozdzielczością, może być trudne i wiąże się ze sporymi nakładami finansowymi [4]. Idealną sytuacją, z punktu widzenia przygotowania prognozy dla operatora SEE, jest minimalizacja nakładów pracy oraz (bardzo często) minimalizacja trudności i kosztów pozyskiwania danych stanowiących zestaw zmiennych objaśniających. Minimalizacja nakładów pracy, obsługi danych w postaci zmiennych objaśniających i kosztów ich pozyskania możliwa jest dzięki:
- zautomatyzowanym procesom typowania i rankingowania najkorzystniejszych zmiennych dla danego procesu (przy zastosowaniu pakietów statystycznych)
- wyborowi najlepszych zmiennych objaśniających z danej grupy pozyskiwanych danych – dzięki ich testowaniu przez różne metody statystyczne.

Jedne z najwyższych kosztów generuje pozyskanie wartości zmiennych objaśniających, będących pomiarami parametrów meteorologicznych (dane historyczne oraz prognozy), które są wykorzystywane do potrzeb gospodarczych i przemysłowych. Parametry meteorologiczne wynikające z położenia geograficznego Polski [5, 6] w sposób znaczący wpływają na obciążenie KSE.
W zależności od długości analizowanego przedziału czasu (szeregu czasowego) w procesie budowy modelu prognostycznego, oprócz doboru zmiennych objaśniających, konieczny jest także optymalny wybór metody prognostycznej [7]. Wybór tych wielkości w dużym stopniu rzutuje na uzyskane wyniki i w dużej mierze zależy od wiedzy i doświadczenia prognosty. W procesie budowy modelu prognostycznego ponowny wybór zmiennych objaśniających i/lub metody prognozowania w szczególności zachodzi w sytuacji, gdy opracowany wcześniej model nie daje satysfakcjonujących wyników [8].
Przeprowadzone symulacje, których wybrane wyniki zamieszczono w niniejszej publikacji, dotyczyły próby wykorzystania zautomatyzowanej metody doboru najlepszych zmiennych objaśniających. Przez najlepsze zmienne objaśniające rozumie się zmienne, które jak najprecyzyjniej opisują zmienną objaśnianą. Najlepsze zmienne objaśniające typowano w powiązaniu z danymi historycznymi szczytowej wartości zapotrzebowania mocy 15-minutowej w ciągu doby KSE, za pomocą wybranych metod statystycznych. W kolejnym kroku przeliczono potencjalne koszty ich całorocznego zakupu dla rozdzielczości godzinowej obciążenia dobowego. Pozwoliło to na ocenę możliwości ograniczenia liczby parametrów, dla których kupowane mogą być prognozy z wyprzedzeniem tygodniowym i z rozdzielczością godzinową. Idealnym stanem jest sytuacja, w której za pomocą jednego parametru wejściowego (zmiennej objaśniającej) możliwe jest opracowanie prognozy analizowanego parametru. Na obciążenie KSE jednakże wpływa więcej niż jeden parametr, dlatego też wydaje się, że każde ograniczenie nakładów finansowych i pracy na przygotowanie prognozy końcowej może być korzystne dla prognosty. Zaprezentowane podejście do optymalizacji liczby rozpatrywanych parametrów wejściowych modelu prognostycznego może być przydatne dla trzech grup użytkowników:
- doświadczonych badaczy prognostów, których modele (metody) prognostyczne uodporniły się na obserwowane w przyrodzie zmiany (reagują z opóźnieniem na dynamikę ich zachowań)
- początkujących badaczy prognostów, którzy posiadają podstawową wiedzę z zakresu przygotowywania danych na potrzeby prognoz i prognozowania, którzy działają przy ograniczonych zasobach czasowych
- osób zarządzających kosztami pozyskiwania danych wejściowych do opracowywania prognoz.

Do symulacji jako zmienną objaśnianą (prognozowaną) przyjęto maksymalną (szczytową) wartość mocy 15-minutowej zapotrzebowania mocy w ciągu doby KSE [9]. Model prognostyczny zasilono na wejściu danymi historycznymi o kształtowaniu się poszczególnych parametrów, by następnie – po wytypowaniu liczby optymalnych parametrów wejściowych – zasymulować roczne koszty zakupu prognoz wytypowanych parametrów. Jako najkorzystniejsze rozwiązanie zadanego problemu założono minimalizację nakładów pracy i nakładów finansowych związanych z pozyskiwaniem danych wejściowych do procesu opracowania prognoz.

This is a supporting translation of the original text published in this issue of "Acta Energetica" on pages 37–42. When referring to the article please refer to the original text.

PL

2. Zmienne objaśniające meteorologiczne (zmienne zewnętrzne)

Na potrzeby niniejszej publikacji jako zbiór zmiennych objaśniających meteorologicznych wybrano pomiary meteorologiczne z jednej z lokalizacji na południu Polski, które uznano w dużym przybliżeniu jako odzwierciedlające średnie warunki meteorologiczne dla całego KSE. Lokalizacja ta nie stanowi ani bieguna zimna (Suwałki), ani bieguna ciepła (Wrocław/Legnica), a porównanie obserwacji z tej stacji z danymi historycznymi kilku innych portali pogodowych pozwalają oszacować, że jest ona w dużym przybliżeniu zgrubną średnią arytmetyczną (ok. ±1°C) z obu biegunów temperatury. Pierwszy zestaw zmiennych (najliczniejszy) składał się z 14 następujących parametrów meteorologicznych, które mogą być przedmiotem zakupu od specjalistycznych podmiotów (tab. 1).

Zmienne objaśniające meteorologiczne	Jedn.	Kod
Temperatura otoczenia maksymalna	°C	Zm6
Temperatura otoczenia minimalna	°C	Zm7
Opady deszczu	mm	Zm8
Prędkość wiatru średnia	km/h	Zm9
Prędkość wiatru średnia ekspercka	km/h	Zm10
Prędkość wiatru maksymalna	km/h	Zm11
Ciśnienie atmosferyczne	hPa	Zm12
Liczba stopniodni grzewczych	°C dzień	Zm13
Liczba stopniodni chłodniczych	°C dzień	Zm14
Liczba godzin słonecznych	–	Zm15
Energia słoneczna	W/m²	Zm16
Poziom promieniowania UV	–	Zm17
Temperatura punktu rosy	°C	Zm18
Temperatura mokrego termometru	°C	Zm19

Tab. 1. Przykładowe parametry meteorologiczne stanowiące zmienne objaśniające i mogące być przedmiotem zakupu

Prezentację maksymalnych i minimalnych dobowych wartości temperatury otoczenia ze wspomnianej lokalizacji za okres od początku stycznia 2010 roku do końca grudnia 2014 roku zamieszczono na rys. 1.

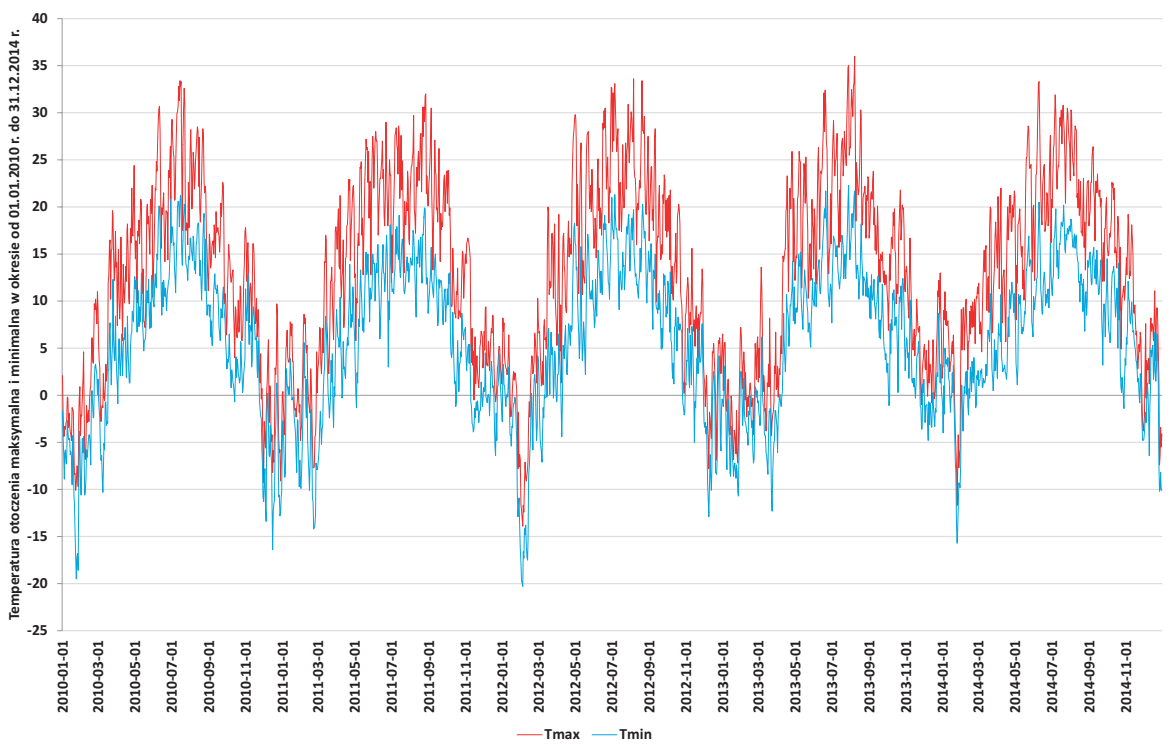
Do powyższego zestawu dołączono dwie zmienne objaśniające, które uzupełniają powyższy zbiór. Pierwsza z nich (Zm2) zawiera w postaci zakodowanej informację o dacie dokonanego pomiaru (rok/miesiąc/dzień) z rozróżnieniem kolejnych dni tygodnia oraz z uwzględnieniem podziału na dzień nieświąteczny i świąteczny. Druga ze wspomnianych zmiennych (Zm3) ma postać czasową niezakodowaną, w której zawarto informację o czasie (z rozdzielczością 15-minutową) wystąpienia szczytowej wartości mocy 15-minutowej obciążenia KSE oraz o generacji wiatrowej w KSE [9] w każdej dobie analizowanego szeregu czasowego.

3. Zmienne objaśniające (zmienne wewnętrzne)

Zmienną objaśnianą, w części dotyczącej typowania zmiennych objaśniających, była wspomniana szczytowa wartość mocy 15-minutowej dobowego obciążenia KSE. Późniejsze symulacje przeprowadzono dla przypadku prognozowania średniego obciążenia godzinowego KSE dla całego roku kalendarzowego (52 tygodnie).

Dodatkowo, w celu uwydatnienia wpływu warunków wiatrowych na obciążenie KSE, przetestowano zmienne Zm20–23 (tab. 2), a także podjęto próbę oceny przydatności zakodowanej informacji o cyklu księżyca (Zm5) do wyjaśniania zmiennej objaśnianej. Wykaz zmiennych objaśniających wewnętrznych przedstawiono w tab. 2.

Poszczególne zmienne objaśniające stanowią ciągi danych historycznych pobranych



Rys. 1. Maksymalne i minimalne wartości dobowe temperatury otoczenia za okres pięciu lat kalendarzowych

This is a supporting translation of the original text published in this issue of "Acta Energetica" on pages 37–42. When referring to the article please refer to the original text.

PL

raz na dobę i obejmują okres pięciu lat, tj. 1 stycznia 2010 – 31 grudnia 2014.
Do metod zautomatyzowanych pozwalających na typowanie zmiennych objaśniających zaliczono m.in. (z odpowiednim kryterium) metody przedstawione w tab. 3. Do innych metod typowania zmiennych objaśniających zaliczono metody wymienione w tab. 4.
Po zastosowaniu opisanego powyżej podejścia uzyskano następujące zestawy zmiennych objaśniających dla poszczególnych metod – tab. 5.

4. Analiza potencjalnych kosztów pozyskania prognoz danych wejściowych dla prognoz dobowego obciążenia KSE wszystkich dób w roku z rozdzielczością godzinową

Zestawienie wstępnie wytypowanych zestawów zmiennych objaśniających zaprezentowano w tab. 6, a ich interpretację graficzną przedstawiono na rys. 2.
Analiza danych z rys. 2 wskazuje, że dla zmiennych płatnych (kolor czerwony) 3 spośród 5 metod zautomatyzowanych wytypowało najmniej liczne zestawy zmiennych wejściowych, potrzebnych do opracowania jak najdokładniejszej prognozy szczytowego dobowego obciążenia KSE. Z wymienionej trójki najmniejszą liczbę parametrów wejściowych wytypowała metoda MARS oraz metoda wykorzystująca współczynnik Pearsona (5 zmiennych). Metoda regresji wielorakiej wytypowała 6 zmiennych wejściowych.
W celu przeprowadzenia symulacji rocznych kosztów zakupu prognoz założono 4 warianty (W1–W4) kosztów pozyskania prognozy na tydzień w przód, z rozdzielczością godzinową dla pojedynczego parametru 100–1000 zł. Na potrzeby tej symulacji założono, że prognoza będzie dotyczyć obciążenia KSE na każdą godzinę doby w roku, bez określania, czy jest to wartość maksymalna, średnia czy minimalna. Zestawienie danych wykorzystanych do symulacji kosztowej przedstawiono w tab. 7 oraz na rys. 3. Dodatkowo w tab. 7 zamieszczono uśrednione arytmetycznie wartości czterech mierników skuteczności prognoz [11]. Zastosowane mierniki to współczynniki: MPE, MAPE, RMSPE, Theila. Uśrednienie dotyczyło błędów prognoz wyrażonych w proc. za okres pięciu lat wstecz przy wykorzystaniu 15 metod prognostycznych wymienionych w tab. 8.
Analiza danych z tab. 7 i rys. 3 wskazuje, że metody M2, M3 i M4 – dzięki wytypowaniu najmniejszej liczby zmiennych płatnych spośród całkowitych liczb zmiennych objaśniających – pozwalają na uzyskanie najniższych uśrednionych kosztów zakupu prognoz tych zmiennych w skali roku (52 tygodni). Dodatkowo metoda M2 pozwala, oprócz minimalizacji kosztów zakupu prognoz zmiennych wejściowych, na uzyskanie najdokładniejszych prognoz *ex-post* za okres 2010–2014. Ponadto metoda M2 umożliwia minimalizację nakładów czasu poświęcanych na budowę modelu prognostycznego, dzięki umożliwieniu badaczowi wglądu w automatyczną ocenę ważności zmiennych wejściowych z podanego zbioru zmiennych. Tym samym metoda M2 wydaje się jednocześnie najskuteczniejsza, najtańsza i najszybsza w warunkach przeprowadzonego eksperymentu symulacyjnego.

Inne zmienne objaśniające	Jedn.	Kod
Udział dobowej szczytowej mocy 15-minutowej w szczycie tygodniowym	%	Zm4
Zakodowana informacja o fazie księżyca	–	Zm5
Maksymalna generacja w farmach wiatrowych	MW	Zm20
Godzina wystąpienia maksymalnej generacji w farmach wiatrowych	–	Zm21
Moc osiągalna w farmach wiatrowych	MW	Zm22
Udział mocy generowanej w farmach wiatrowych do ich mocy zainstalowanej	%	Zm23

Tab. 2. Inne parametry stanowiące zmienne objaśniające [9, 10]

Zautomatyzowane metody typujące zmienne objaśniające	Kryterium	Kod
Regresja wieloraka (metoda klasyczna)	B > ±0,04	M1
Metoda MARS (metoda *data mining*)	ranking predyktorów	M2
Obliczanie współczynnika Pearsona (metoda klasyczna)	>0,47	M4
Metoda szybka C&RT (metoda *data mining*)	ranking predyktorów	M8
Dobór i eliminacja zmiennych (metoda *data mining*)	ranga zmiennej	M10

Tab. 3. Wybrane zautomatyzowane metody typujące zmienne objaśniające

Inne metody typujące zmienne objaśniające	Kryterium	Kod
Wybór zmiennych objaśniających postrzeganych jako posiadające znaczący wpływ na obciążenie KSE	autorskie	M3
Wybór wszystkich posiadanych zmiennych objaśniających	brak	M5
Wybór wszystkich spośród najlepszych zmiennych objaśniających z metod M1–M4	zgodnie z kryt. dla metod M1–M4	M6
Wybór wszystkich spośród najlepszych zmiennych objaśniających oraz ekspercki dobór dodatkowej/dodatkowych zmiennych objaśniających	brak + autorskie	M7
Regresja wieloraka przeprowadzana iteracyjnie osobno dla każdej zmiennej objaśnianej	B > ±0,1	M9

Tab. 4. Inne metody typujące zmienne objaśniające

Metoda	Zmienne	Liczba zmiennych	Liczba zmiennych płatnych
M1	4, 6–7, 13, 16–17, 19	7	6
M2	3–4, 6–7, 16–17, 20, 22	8	5
M3	11, 13–15, 18	5	5
M4	4, 6–7, 13, 16–17	6	5
M5	1, 3–23	22	14
M6	3–4, 6–7, 11, 13–20, 22	14	10
M7	3–4, 6–7, 11–20, 22 (dodatkową zm. jest ciśnienie Z12)	15	11
M8	4, 6–7, 12–20, 22	13	10
M9	4, 6–7, 13–14, 16–19, 22–23	11	8
M10	3–4, 6–7, 13–14, 16–19	10	9

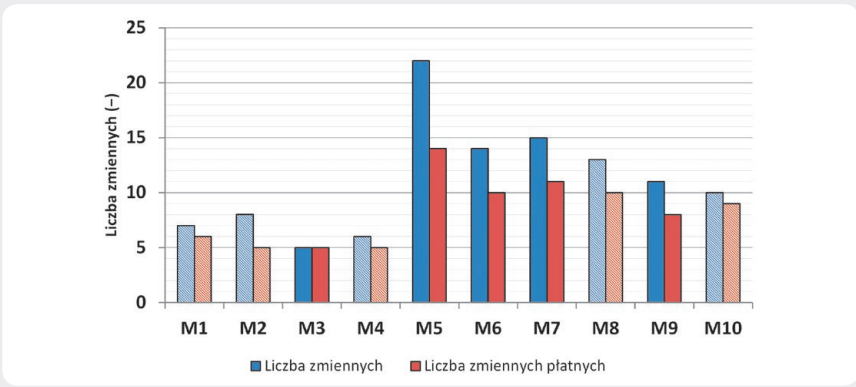
Tab. 5. Wykaz zestawów zmiennych objaśniających z podziałem na całkowitą ich liczbę i liczbę zmiennych płatnych

5. Podsumowanie

Podstawowym celem prowadzonych badań, których wybrane wyniki zostały przedstawione w niniejszej publikacji, było

This is a supporting translation of the original text published in this issue of "Acta Energetica" on pages 37–42. When referring to the article please refer to the original text.

PL



Uwaga! kreskowaniem zaznaczono wyniki uzyskane przez metody zautomatyzowane

Rys. 2. Liczba zmiennych objaśniających w poszczególnych zestawach danych wytypowanych przy użyciu wytypowanych metod doboru zmiennych

Metoda	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Liczba zmiennych	**7**	**8**	5	**6**	22	14	15	**13**	11	**10**
Liczba zmiennych płatnych	**6**	**5**	5	**5**	14	10	11	**10**	8	9

Uwaga! Czcionką pogrubioną zaznaczono metody zautomatyzowane

Tab. 6. Wstępnie wytypowane liczby zmiennych objaśniających z zestawu 22 zmiennych wejściowych (objaśniających)

zautomatyzowanie i wytypowanie optymalnego doboru zestawu zmiennych objaśniających z założeniem minimalizacji nakładów finansowych i nakładów pracy. Mając na uwadze powyższe, wnioskuje się, że zalecany jest wybór metody M1 (regresja wieloraka), M2 (MARS) oraz M4 (współczynnik Pearsona) typowania zestawu zmiennych. Zauważalne jest także, że dla zaprezentowanego zestawu zmiennych płatnych najmniejszą liczbę wymaganych danych wytypowano również w oparciu o badania literaturowe, doświadczenie i wiedzę autorską. Wyniki uzyskane przez metody *data mining* M8 (C&RT) oraz M10 (dobór i eliminacja zmiennych) prawie dwukrotnie przekraczały wyniki uzyskane dla najlepszych metod, dlatego z większą ostrożnością należy podchodzić do typowania przez te dwie metody w przyszłości. Przedstawione podejście wskazuje na jego przydatność do wstępnego optymalizowania przewidywanych kosztów opracowywania prognoz.

Założenie czterech różnych scenariuszy kształtowania się cen prognoz poszczególnych zmiennych objaśniających wskazuje, że M2 (MARS), M3 (wsp. Pearsona) oraz M4 (metoda ręcznego doboru) bez względu na rozkład kosztów zakupu prognoz poszczególnych zmiennych objaśniających będą dawać najniższe średnioroczne koszty zakupu (<200 tys. zł) dla założonych wariantów kosztów zakupu. Powyższe wynika z faktu, że dla każdej z metod wytypowana liczba zmiennych płatnych wynosiła 5. Tym samym wykazana została odporność tych trzech metod na potencjalnie różne koszty zakupu danych wejściowych. Wybór najkorzystniejszej metody spośród trzech najmniej kosztownych, która w przyszłości mogłaby okazać się najkorzystniejsza dla budowania skutecznych prognoz, jest możliwy dzięki nałożeniu na rys. 3 średniej arytmetycznej skuteczności *ex-post* wykonywanych prognoz. Z wyróżnionej wcześniej trójki metod odrzucić należy metodę M3, która uzyskała najmniej korzystny wynik (7,11%) w ocenie skuteczności prognozowania za okres pięciu lat. Z pozostałej dwójki metod wyróżniającym się wynikiem skuteczności prognoz *ex-post* charakteryzuje się metoda M2 (MARS), która jako jedyna pozwoliła na uzyskanie skuteczności prognozowania na poziomie poniżej 3% (2,94%). Metoda ta zapewnia tym samym zarówno minimalizację kosztów zakupu prognoz zmiennych wejściowych, dzięki wyposażeniu badacza prognosty w możliwość zautomatyzowanego typowania zmiennych objaśniających, jak i zapewnia najwyższą skuteczność prognozowania *ex-post* spośród wytypowanych w tej publikacji metod. Tym samym należy uznać, że wykorzystanie metody MARS pozwala na minimalizację czasu potrzebnego na przygotowanie danych wejściowych do modelu prognostycznego, minimalizację kosztów pozyskiwania prognoz tych danych oraz zapewnia (w realiach zastosowanego podejścia) największą dokładność uzyskiwanych za jej pomocą prognoz.

Konieczna jest weryfikacja skuteczności opracowanych zestawów danych (zmiennych) objaśniających (wejściowych) dla prognozowania *ex-ante* krótkoterminowego dobowego obciążenia KSE z rozdzielczością

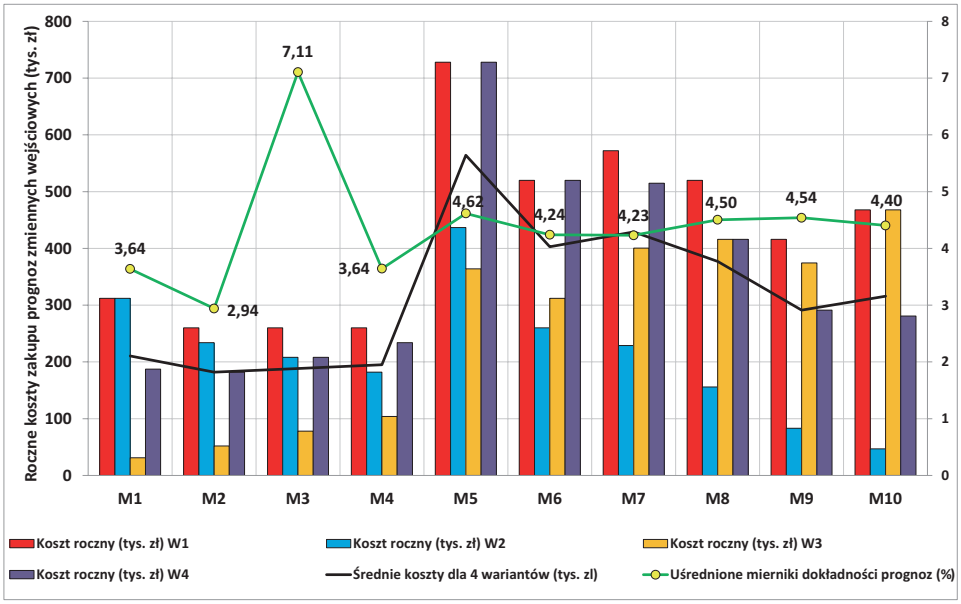
Metoda	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Liczba zmiennych płatnych (–)	6	5	5	5	14	10	11	10	8	9
Liczba tygodni (–)	52	52	52	52	52	52	52	52	52	52
Koszt tygodniowy za prognozę zmiennej wejściowej (tys. zł) W1	1	1	1	1	1	1	1	1	1	1
Koszt tygodniowy za prognozę zmiennej wejściowej (tys. zł) W2	1	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1
Koszt tygodniowy za prognozę zmiennej wejściowej (tys. zł) W3	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Koszt tygodniowy za prognozę zmiennej wejściowej (tys. zł) W4	0,6	0,7	0,8	0,9	1	1	0,9	0,8	0,7	0,6
Koszt roczny (tys. zł)	312	260	260	260	728	520	572	520	416	468
Koszt roczny (tys. zł) W1	312	234	208	182	437	260	229	156	83	47
Koszt roczny (tys. zł) W2	31	52	78	104	364	312	400	416	374	468
Koszt roczny (tys. zł) W3	187	182	208	234	728	520	515	416	291	281
Koszt roczny (tys. zł) W4	211	182	189	195	564	403	429	377	291	316
Średni roczny koszt z wariantów W1–W4	**211**	**182**	**189**	**195**	**564**	**403**	**429**	**377**	**291**	**316**
Uśrednione mierniki oceny dokładności prognoz (%)	**3,64**	**2,94**	**7,11**	**3,64**	**4,62**	**4,24**	**4,23**	**4,50**	**4,54**	**4,40**

Uwaga! Czcionką pogrubioną zaznaczono metody zautomatyzowane

Tab. 7. Symulacja rocznych kosztów zakupu prognoz zmiennych wejściowych na potrzeby krótkoterminowej prognozy dobowego obciążenia KSE na tle uśrednionych mierników skuteczności prognoz *ex-post*

This is a supporting translation of the original text published in this issue of "Acta Energetica" on pages 37–42. When referring to the article please refer to the original text.

PL



Rys. 3. Symulacja rocznych kosztów zakupu prognoz zmiennych wejściowych na potrzeby krótkoterminowej prognozy dobowego obciążenia KSE na tle uśrednionych mierników skuteczności prognoz *ex-post*

Metoda	Skrót	Kod
Wielozmienna regresja adaptacyjna z użyciem funkcji sklejanych	MARS	S1
Ogólne modele drzew klasyfikacyjnych i regresyjnych (wersja standardowa)	C&RT std.	S2
Ogólne modele drzew klasyfikacyjnych i regresyjnych (wersja z układami)	C&RT z ukł.	S3
Automatyczny detektor interakcji za pomocą chi-kwadrat (wersja standardowa)	CHAID std.	S4
Automatyczny detektor interakcji za pomocą chi-kwadrat (wersja z układami)	CHAID z ukł.	S5
Drzewa interakcyjne regresyjne (wersja interaktywna)	C&RT inter.	S6
Drzewa interakcyjne regresyjne (wersja interaktywna)	CHAID inter.	S7
Drzewa interakcyjne regresyjne (wersja wyczerpująca)	CHAID wycz.	S8
Uogólnione modele addytywne z użyciem funkcji wiążącej identycznościowej	UMA iden.	S9
Uogólnione modele addytywne z użyciem funkcji wiążącej logarytmicznej	UMA log.	S10
Regresja wieloraka	Regr. Wlrk.	S11
Ogólne modele liniowe i nieliniowe	OMLN	S12
Ogólne modele regresji	OMR	S13
Modele najmniejszych kwadratów (cząstkowa)	MNKC	S14
Sztuczne sieci neuronowe	SSN	S15

Uwaga! Czcionką pogrubioną zaznaczono metody zautomatyzowane

Tab. 8. Metody statystyczne wykorzystane do badań symulacyjnych

godzinową. Weryfikacja taka powinna obejmować zbiór metod klasycznych oraz zgłębiania danych (*data mining*).

Należy zauważyć ciekawą zależność, że wykonane badania i analiza porównawcza wyników (przeprowadzona na podstawie tab. 6 i 7) wskazują, że najkorzystniej będzie się skupić na metodzie MARS, która z grupy 8 danych (zmiennych) wejściowych wytypowała jedynie 5 zmiennych związanych z kosztami zakupu [11]. Metoda ta należy do grupy metod *data mining* i oprócz analizy statystycznej oferuje szybką i zautomatyzowaną drogę do uzyskania najkorzystniejszego zestawu zmiennych objaśniających. Warto podkreślić również, że uzyskane wyniki są najkorzystniejsze dla rozpatrywanego 5-letniego zbioru danych historycznych. Należy zaznaczyć jednakże na obecnym etapie, że metoda MARS na etapie uczenia (prognozowanie w trybie *ex-post*) jest podatna na przeuczenie i w zderzeniu z realnym prognozowaniem może dawać mniej obiecujące wyniki.

Cennym doświadczeniem mogłoby być uwzględnienie jako zmiennych objaśniających danych historycznych o cenach energii elektrycznej na Rynku Bilansującym, TGE lub przebiegów notowań ciągłych towarów giełdowych, np. ropy Brent w ramach rynków *intraday* i *day ahead* [12].

Bibliografia

1. Bartodziej G., Tomaszewski M., Polityka energetyczna i bezpieczeństwo energetyczne – wydanie II, Wydawnictwo „Nowa Energia", Racibórz 2008.
2. Weron A., Weron R., Giełda Energii: strategie zarządzania ryzykiem, Wydawnictwo CIRE, Centrum Metod Stochastycznych im. Hugona Steinhausa, Politechnika Wrocławska, Wrocław 2000.
3. Klóska R., Hundert M., Czyżycki R., Wybrane zagadnienia z prognozowania, Wydawnictwo Economicus, Szczecin 2007.
4. Zeliaś A., Pawełek B., Wanat S., Prognozowanie ekonomiczne. Teoria, przykłady, zadania, Wydawnictwo Naukowe PWN, Warszawa 2003.
5. Prognozowanie w elektroenergetyce. Zagadnienia wybrane, red. I. Dobrzyńska, Wydawnictwo Politechniki Śląskiej, Częstochowa 2012.
6. Kopecki K. i in., Analiza i prognoza obciążeń elektroenergetycznych, Wydawnictwa Naukowo-Techniczne, Warszawa 1971.
7. Witkowska D., Podstawy ekonometrii i teorii prognozowania. Podręcznik z przykładami i zadaniami, Oficyna Ekonomiczna, Kraków 2005.
8. Kot S.M., Jakubowski J., Sokołowski A., Statystyka. Podręcznik dla studiów ekonomicznych, Centrum Doradztwa i Informacji, Warszawa 2007.
9. www.pse.pl.

This is a supporting translation of the original text published in this issue of "Acta Energetica" on pages 37–42. When referring to the article please refer to the original text.

PL

10. Pieczarko R., Sołtysik M., Analiza wpływu generacji źródeł wiatrowych na poziom kształtowania się cen energii elektrycznej na rynku SPOT, Konferencja Naukowa „Prognozowanie w elektroenergetyce", Podlesice 2016.
11. Czapaj R., Rzepka P., Szablicki M., Typowanie zmiennych objaśniających przy wykorzystaniu zautomatyzowanych metod statystycznych jako sposób optymalizacji wyboru metody estymacji szczytowego dobowego obciążenia KSE, Konferencja Naukowa „Prognozowanie w elektroenergetyce", Podlesice 2016.
12. Kozakiewicz M. i in., Zastosowanie ekonometrycznych modeli prognostycznych w transakcjach proprietary trading, Konferencja Naukowa „Rynek energii elektrycznej", Kazimierz Dolny 2015.

Rafał Czapaj
mgr inż.
PSE Innowacje sp. z o.o.
e-mail: rafal.czapaj@pse.pl
Absolwent Wydziału Elektrycznego Politechniki Śląskiej (2003). Z Grupą Kapitałową PSE SA związany jest od stycznia 2005 roku. W latach 2005–2011 pracował w EPC SA i zajmował się tematyką rynku energii elektrycznej oraz analiz techniczno-ekonomicznych. Od 2011 roku pracuje w PSE Innowacje sp. z o.o. (wcześniej CATA), zajmując się tą samą tematyką.