

Prognozowanie krótkoterminowe zapotrzebowania na energię elektryczną w KSE z wykorzystaniem metody MARSplines

Streszczenie. W tekście przedstawiono wyniki zastosowania metody MARSplines, należącej do szerokiej grupy metod Data Mining, do prognozowania zapotrzebowania na moc elektryczną w Krajowym Systemie Elektroenergetycznym. Przedstawiono zmienne objaśniające i ich powiązanie ze zmienną prognozowaną. Zamieszczono ponadto obszerną analizę statystyczną tych zmiennych. Na podstawie przeprowadzonych symulacji w trybie *ex post* i *ex ante* sformułowano wnioski dotyczące samej metody oraz wnioski wynikające z czynności progностycznych.

Abstract. The article presents the results obtained from applying the MARSplines method, which belongs to a broad group of Data Mining methods, to forecast the electric power demand in the Polish National Power System. Furthermore, the relationship between explanatory variables and the forecasted variable is examined through an extensive statistical analysis. Based on the (*ex-post* and *ex-ante*) simulations results a number of conclusions are drawn regarding the method itself and the accuracy of its predictions. (**Short-Term National Power System Electricity Demand Forecasting using the MARSplines Method**).

Słowa kluczowe: metoda MARSplines, zmienne objaśniające, analiza statystyczna, prognozowanie, KSE, zapotrzebowanie na moc.

Keywords: MARSplines Method, Explanatory Variable, Statistical Analysis, Forecasting, National Power System, Power Load.

Wstęp

Jedną z metod należących do grupy metod zgłębiania danych (ang. *Data Mining*) jest metoda *Multivariate Adaptive Regression Splines* (MARSplines). Jest to wielozmienna adaptacyjna metoda regresyjna wykorzystująca tzw. Splajny, czyli funkcje sklepane. Metoda została zaproponowana w 1991 r. przez Jerome'a H. Friedman'a [1]. Algorytm początkowo był zaprojektowany dla zbiorów danych, które nie przekraczały 1000 obserwacji i dotyczyły maksymalnie 20 zmiennych objaśniających [1]. Obecnie metoda MARSplines oznaczona jest znakiem towarowym, a właścicielem jej licencji jest firma Salford Systems. Podstawowym celem artykułu jest próba weryfikacji głównych wad analizowanej metody na przykładzie krótkoterminowej prognozy zapotrzebowania na moc w Krajowym Systemie Elektroenergetycznym.

Do głównych wad opisywanych w literaturze przedmiotu można zaliczyć m.in. [2, 3, 4]:

- wrażliwość na obserwacje odstające – zastosowano 4 szeregi zawierające obserwacje odstające;
- znaczący stopień arbitralności w procesie doboru zmiennych objaśniających (w szczególności w przypadku współliniowości tych zmiennych) – przebadano jakość prognoz *ex post* dla wszystkich zmiennych oraz dla zmiennych typowanych przez analizowaną metodą;
- znaczące zagrożenie nadmiernym dopasowaniem do danych wejściowych co wynika z zastosowania metody rekurencyjnego podziału przestrzeni cech (ang. *Recursive Partitioning*) – zastosowano szeregi czasowe o kilkuletniej historii dla każdej doby.

Omawiana metoda należy do grupy metod stosujących uczenie *nieukierunkowane*. W takim podejściu poszukiwany jest model najlepiej pasujący do danych, bez uwzględnienia mechanizmu jaki te dane wygenerował, a tworzony model (algorytm) nie ma służyć opisowi relacji przyczynowych, a jedynie prognozowaniu lub rozpoznawaniu [5] (typowe zagadnienia metod z grup *Data Mining* [6,7,8,9,10]). Omawiana metoda jest metodą ukierunkowaną (ang. *Supervised Learning*).

Zmienne objaśniające i ich opis

Jako zbiór zmiennych objaśniających wybrano pomiary meteorologiczne pozyskane z lokalizacji na środkowym południu Polski, które można uznać jako „reprezentatywne” dla całego KSE. Wymieniona lokalizacja nie stanowi „bieguna zimna” (Suwałki) ani nie stanowi „bieguna ciepła”

(Wrocław), a wieloletnie obserwacje pozwalają oszacować, że jest ona w dużym przybliżeniu zgrubną średnią [11, 12] (poniżej $\pm 1^\circ\text{C}$) z obu biegunów temperatury [13, 14]. Pierwszy zestaw zmiennych (najliczniejszy) składa się z 7 parametrów meteorologicznych: temperatury otoczenia [$^\circ\text{C}$] – ZO1; wilgotności powietrza [%] – ZO2; porywów wiatru [km/h] – ZO3; prędkości wiatru [km/h] – ZO4; kierunku wiatru [-] – ZO5; temperatury odczuwalnej [$^\circ\text{C}$] – ZO6; temperatury punktu Rosy [$^\circ\text{C}$] – ZO7.

Dane źródłowe dotyczące zapotrzebowania na moc elektryczną w Krajowym Systemie Elektroenergetycznym pozyskano z publikacji PSE S.A. [15] dla każdej doby analizowanego szeregu czasowego w latach 2009-2015. Zmienną objaśnianą była wyłącznie średnia wartość zapotrzebowania na moc (w pierwszej godzinie doby). Założono, że tak dobrany zestaw zmiennych objaśniających będzie stanowił znaczące wyzwanie dla analizowanej metody ze względu na ich zróżnicowanie i niejednorodność oraz charakter opisany parametrami statystycznymi opisanymi w dalszej części niniejszej pracy.

Statystyki badanych zmiennych

Wpływ każdej zmiennej objaśniającej na zmienną objaśnianą opisano przy pomocy macierzy korelacji (tab. 1).

Tabela 1. Macierz korelacji

Zm.	Zap.	ZO1	ZO2	ZO3	ZO4	ZO5	ZO6	ZO7
Zap.	1,000	-0,538	0,210	-0,512	0,012	0,055	-0,002	-0,192
ZO1	-	1,000	-0,356	0,960	-0,032	-0,054	-0,025	0,239
ZO2	-	-	1,000	-0,088	0,015	0,018	0,184	0,102
ZO3	-	-	-	1,000	-0,027	-0,049	0,027	0,284
ZO4	-	-	-	-	1,000	0,970	-0,077	0,043
ZO5	-	-	-	-	-	1,000	-0,066	0,020
ZO6	-	-	-	-	-	-	1,000	0,065
ZO7	-	-	-	-	-	-	-	1,000

Legenda: Pogrubieniem oznaczono korelacje o poziomie istotności $p < 0,05$. Kursywą oznaczono wartości najwyższe dla korelacji dodatnich i ujemnych.

Przedstawiona macierz korelacji wskazuje, że zmienna objaśniana dodatnio najmocniej koreluje ze zmienną objaśniającą ZO2 (+0,210), natomiast ujemnie najmocniej koreluje ze zmienną objaśniającą ZO1 (-0,538). Niestety w obu przypadkach współczynnik ten oznaczają niską korelację ze zmienną objaśnianą, co stanowi dodatkowe utrudnienie w odniesieniu do zadania progностycznego postawionego w tym artykule. W celu szerszego poznania sposobu kształtowania się poszczególnych zmiennych

objaśniających przedstawiono poszerzony zakres statystyk opisowych. Badaniom poddano 2 556 przypadków, których statystyki opisowe zaprezentowano w tabeli 2 (cz.1 i cz.2).

Tabela 2. Statystyki opisowe dla zmiennej objaśnianej oraz zmiennych objaśniających – cz.1

Zmienna/Miara	Zap.	ZO1	ZO2	ZO3
Srednia	15 321,48	8,21	86,03	5,82
Statystyka Grubbsa	3,42	3,27	4,31	3,51
Poziom p	1,00	1,00	0,04	1,00
Mediana	15 335,26	8,70	89,20	6,15
Minimum	10 678,70	-19,30	34,30	-21,30
Maksimum	20 037,78	28,70	100,00	21,60
Dolny kwartyl	14 381,56	2,00	79,50	0,30
Górny kwartyl	16 305,20	14,50	95,20	11,95
Percentyl 10,0	13 441,68	-2,70	68,30	-3,90
Percentyl 90,0	17 099,40	19,00	98,20	15,70
Rozstęp kwartyla	1 923,64	12,50	15,70	11,65
Wariancja	1 906 356,62	70,58	143,89	59,74
Odch. std.	1 380,71	8,40	12,00	7,73
P. ufnosci odch. std.*	1 343,87	8,18	11,68	7,52
P. ufnosci odch. std.**	1 419,64	8,64	12,33	7,95
Współczynnik zmienności	9,01	102,35	13,94	132,79
Standardowy błąd	27,31	0,17	0,24	0,15
Skośność	-0,13	-0,20	-1,16	-0,35
Kurtoza	-0,21	-0,38	1,07	-0,29

Legenda: * - P. ufnosci odch. std. – Poziom ufnosci odchylenia standardowego (-95%); ** - P. ufnosci odch. std. – Poziom ufnosci odchylenia standardowego (+95%).

Tabela 2. Statystyki opisowe dla zmiennej objaśnianej oraz zmiennych objaśniających – cz.2

Zmienna/Miara	ZO4	ZO5	ZO6	ZO7
Srednia	2,51	5,24	183,20	27,50
Statystyka Grubbsa	16,02	15,18	2,06	6,82
Poziom p	0,00	0,00	1,00	0,00
Mediana	1,60	3,80	183,70	19,50
Minimum	0,00	0,00	0,00	0,00
Maksimum	108,30	127,70	360,00	234,40
Dolny kwartyl	0,50	2,20	112,00	5,90
Górny kwartyl	3,10	6,50	251,50	36,80
Percentyl 10,0	0,10	1,00	68,00	0,40
Percentyl 90,0	4,75	9,70	315,00	66,40
Rozstęp kwartyla	2,60	4,30	139,50	30,90
Wariancja	43,60	65,05	7 893,48	921,00
Odch. std.	6,60	8,07	88,85	30,35
P. ufnosci odch. std.*	6,43	7,85	86,47	29,54
P. ufnosci odch. std.**	6,79	8,29	91,35	31,20
Współczynnik zmienności	263,53	153,99	48,50	110,36
Standardowy błąd	0,13	0,16	1,76	0,60
Skośność	12,61	10,35	0,09	2,41
Kurtoza	175,31	132,22	-0,98	9,31

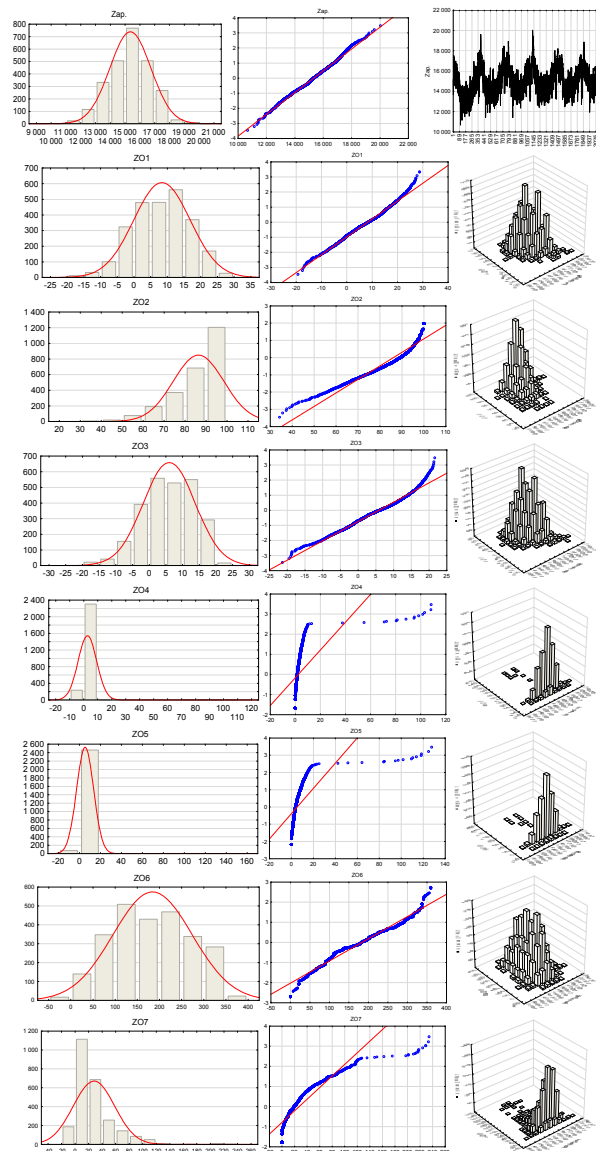
Legenda: * - P. ufnosci odch. std. – Poziom ufnosci odchylenia standardowego (-95%); ** - P. ufnosci odch. std. – Poziom ufnosci odchylenia standardowego (+95%).

Szereg czasowy zmiennej objaśnianej charakteryzuje się sezonowością i silną okresowością oraz nie ma rozkładu normalnego (tab. 3). Nieduża wartość współczynnika zmienności wskazuje, że uzyskanie wysokiej skuteczności prognoz nie powinno nastręczać dużych trudności. Współczynniki korelacji liniowej Pearsona posiadają niskie wartości. Nieznaczna ujemna wartość kurtozy wskazuje na mniejszą koncentrację wartości wokół wartości średniej (rozkład jest bardziej spłaszczony od normalnego). Współczynnik skośności o niewielkiej wartości ujemnej wskazuje na asymetrię lewostronną szeregu (więcej wartości mniejszych od wartości średniej).

Rozkłady prawdopodobieństwa i wykresy normalności

Wizualną reprezentację rozkładów prawdopodobieństwa oraz normalności poszczególnych szeregów czasowych przedstawia rysunek 1. Do niekorzystnych rozkładów prawdopodobieństwa zaliczyć należy rozkłady dotyczące wszystkich zmiennych objaśniających. Najbardziej

niekorzystne wykresy normalności odnoszą się do zmiennych objaśniających ZO3, 5, 6 oraz 7 (rys. 1). Dodatkowo zmienną objaśnianą zaprezentowano na tym rysunku przy pomocy wykresu liniowego (za okres lat 2009-2014), a w przypadku zmiennych objaśnianych zilustrowano w przestrzeni trójwymiarowej ich porównanie ze zmienną objaśnianą. Wykresy trójwymiarowe przedstawione na rysunku 1 wskazują na duże zróżnicowanie i stanowią wyzwanie dla analizowanej metody i będą służyć weryfikacji głównych wad metody przedstawionych na wstępie.



Rys.1. Histogramy oraz wykresy normalności zmiennej zależnej oraz zmiennych objaśniających

Ze względu na niedoskonałości wynikające z możliwości i skuteczności wizualnej oceny normalności zmiennych poddawanych dalszym badaniom przeprowadzono weryfikację hipotezy zerowej mówiącej o normalności każdego z szeregów czasowych. Dodatkowo przeprowadzono testy losowości serii, testy występowania obserwacji odstających. Testy weryfikacji normalności każdego z szeregów czasowych wskazały konieczność odrzucenia hipotezy zerowej. Testy losowości potwierdziły słuszność hipotezy zerowej jedynie dla zmiennych ZO6 oraz ZO7. Obserwacje odstające na drodze testowej stwierdzono dla zmiennych ZO2, ZO4, ZO5 oraz ZO7 (odpowiednio 1, 16, 21 oraz 19 obserwacji odstających).

Tabela 3. Wyniki testów statystycznych dla zmiennej objaśnianej oraz zmiennych objaśniających

Zm./p	Zap.	ZO1	ZO2	ZO3	ZO4	ZO5	ZO6	ZO7
TN S-W	x	x	x	x	x	X	x	x
TN Lillfrs.	x	x	x	x	x	X	x	x
TL W-Ws	x	x	x	x	x	X	x	x
TL W-Wm	x	x	x	x	x	X	x	x
TL M-K	x	x	x	x	x	X	0,5018	0,2333
TL Bartelsa	x	x	x	x	x	X	x	x
TOO Grubbs	b	b	1	b	16	21	b	19
TN S-W R	x	x	x	x	x	X	x	x
TN Lillfrs. R	x	x	x	x	x	X	x	x
T Wlccx	x	x	x	x	x	X	x	x

Legenda: TN S-W – test Shapiro-Wilka normalności ($\alpha=0,05$); TN Lillfrs. – test Lillieforsa normalności ($\alpha = 0,05$); TL W-Ws - test Walda-Wolfowitza losowości serii dla średniej ($\alpha = 0,05$); TL W-Wm - test Walda-Wolfowitza losowości serii dla mediany ($\alpha = 0,05$); TL M-K - test Manna-Kendalla losowości ($\alpha = 0,05$); TL Bartelsa – test Bartelsa losowości rang ($\alpha = 0,05$); TOO – test Grubbsa występowania obserwacji odstających; TN S-W R – test Shapiro-Wilka normalności zmiennej Różnica ($\alpha=0,05$); TN Lillfrs. R. – test Lillieforsa normalności zmiennej Różnica ($\alpha = 0,05$); T Wlccx – test Wilcoxon dla par obserwacji; x – wynik testu odrzuca hipotezę zerową o zachodzeniu przedmiotu hipotezy; b – brak;

Dostępne opcje i parametry metody MARSplines

Metoda MARSplines domyślnie przyjmuje następujące wartości parametrów (tab. 4).

Tabela 4. Wstępne parametry i ustawienie dostępne dla metody MARSplines w pakiecie STATISTICA

Parametr	Wartość/ustawienie
Zmienne ilościowe zależne	Zapotrzebowanie mocy
Zmienne jakościowe zależne	n.d.
Predyktory ilościowe	ZO1-ZO7
Predyktory jakościowe	n.d.
Maksymalna liczba funkcji bazowych	21
Rząd interakcji	1
Kara	2
Próg	0,0005
Usuwanie nieistotnych funkcji bazowych	TAK
Usuwanie braku danych	Przypadkami
Liczba czynników	14
Liczba funkcji bazowych	13

Legenda: n.d. – nie dotyczy.

Zgodnie z informacją na wstępie niniejszej pracy wskazano, że parametrem prognozowanym (zmienną ilościową zależną) jest średnie godzinowe zapotrzebowanie mocy w pierwszej godzinie doby. Jako predyktory ilościowe wybrano wszystkie zmienne meteorologiczne (zmienne objaśniające – predyktory ilościowe).

Drugą z dostępnych metod usuwania brakujących danych jest metoda zastępowania średnią, której wpływ również oceniono na etapie opracowywania modelu prognostycznego. Badania wstępne przeprowadzono zgodnie z danymi i informacjami zawartymi w powyższej tabeli (Opcje specyfikacji modelu). Dodatkowym nurtem badań wstępnych była ocena potencjalnego wpływu standaryzacji wartości zmiennych objaśniających. W analizowanej metodzie nie odnotowano jednakże takiego wpływu. Metodą krokową przebadano wpływ dostępnych opcji i wartości parametrów na prognozy *ex post* dla okresu 6 pełnych lat (2009-2014). Wskaźnikiem wykorzystanym do oceny jakości prognoz *ex post* był błąd MAPE wyrażony w %. Dla pełnego modelu uwzględniającego wszystkie zmienne objaśniające metoda MARSplines oceniła, że zmienne objaśniające ZO1, ZO2, ZO5 oraz ZO7 w największym stopniu wpływają na stopień objaśniania zmiennej prognozowanej (odpowiednio 2, 4, 3, 2). Siła wpływu pozostałych zmiennych została przez omawianą metodą oceniona na 0, 1 oraz 1 (odpowiednio ZO3, ZO4, ZO7). Każda z symulacji skutkowała stworzeniem nowego, bardziej złożonego, równania opisywanego przez metodę MARSplines. Ze względu na odnotowanie zmniejszania

wartości błędu MAPE przy braku aktywacji opcji usuwania nieistotnych funkcji bazowych (FB), dalsze badania przeprowadzono z takim właśnie ustawieniem. Ogólnie postać funkcji MARSplines (zależność 1) uzyskuje się poprzez sumowanie funkcji bazowych oraz iloczynów tych funkcji z odpowiednimi wagami, określonych wspólnie oznaczeniem $h_m(X)$ [11, 12]:

$$(1) \quad y = f(x) = \beta_0 + \sum_{m=1}^M \beta_m H_{km}(x_{v(k,m)})$$

W analizowanym przykładzie, przy uwzględnieniu wszystkich zmiennych objaśniających uzyskano funkcję następującej postaci (zależność 2): gdzie: sumowanie odbywa przebiega przez wszystkie M składników modelu, a β_0 i β_m to parametry modelu (podobnie jak węzły t każdej funkcji bazowej, wyznaczone z danych) [4].

$$(2) \quad \begin{aligned} f(x) = & 15\,577,94 + 59,10 * \max(0; ZO1 - 13,80) + 120,50 \\ & * \max(0; 13,80 - ZO1) + 10,98 \\ & * \max(0; ZO2 - 67,20) - 31,87 \\ & * \max(0; 67,20 - ZO2) + 80,13 \\ & * \max(0; ZO5 - 17,80) - 63,46 \\ & * \max(0; 17,80 - ZO5) - 148,12 \\ & * \max(0; ZO4 - 0,00) - 1,89 \\ & * \max(0; ZO7 - 176,20) - 1,59 \\ & * \max(0; 176,20 - ZO7) - 2,50 \\ & * \max(0; ZO8 - 9,80) + 51,47 \\ & * \max(0; ZO5 - 5,00) - 49,10 \\ & * \max(0; ZO2 - 87,20) + 100,64 \\ & * \max(0; ZO2 - 93,70) \end{aligned}$$

Ograniczenie zestawu zmiennych objaśniających do zmiennych posiadających największy wpływ na zmienną objaśnianą (zmienne 1, 2, 5, 7) przyjmuje postać:

$$(3) \quad \begin{aligned} f(x) = & 14\,411,00 + 60,83 * \max(0; ZO1 - 13,80) + 121,76 \\ & * \max(0; 13,80 + ZO1) - 34,50 \\ & * \max(0; 67,20 - ZO2) - 76,77 \\ & * \max(0; ZO5 - 17,80) - 1,65 \\ & * \max(0; ZO7 - 206,00) - 1,04 \\ & * \max(0; 206,00 - ZO7) + 68,39 \\ & * \max(0; ZO5 - 7,20) - 74,54 \\ & * \max(0; ZO2 - 87,20) + 109,92 \\ & * \max(0; ZO2 - 92,80) + 20,17 \\ & * \max(0; ZO2 - 76,40) \end{aligned}$$

Kolejne cztery tabele (tab. 5-8) ukazują wyniki opisywane przy pomocy procentowego błędu MAPE dla szeregów czasowych obejmujących okres lat od 2009 do 2014. Szeregi te stanowiły zbiór uczący dla metody MARSplines. Tabela 5 objęła wyniki badań dla wszystkich zmiennych objaśniających z aktywną funkcją usuwania nieistotnych FB. Tabela 6 ukazuje wyniki bazujące na tym samym zestawie zmiennych objaśniających przy braku usuwania nieistotnych FB. W tabelach 7 i 8 zastosowano takie samo podejście, jednakże dla zmiennych objaśniających 1, 2, 5, 7 wytypowanych przez analizowaną metodę jako najbardziej wpływające na objaśnianie zmiennej prognozowanej. W każdej z omówionych powyżej tabel w wierszach podświetlonych na kolor szary zapisano kolejno parametry ustawień, czyli maksymalną liczbę funkcji bazowych, rząd interakcji, wartość kary. Każdorazowo w przypadku braku wpływu zmian ustawień i parametrów do dalszych badań przeznaczony był model o największej skuteczności w sensie procentowej wartości błędu MAPE.

Tabela 5. Błędy MAPE *ex post* (wszystkie zm. obj.) z usuwaniem nieistotnych FB (MAPE dla ustawień domyślnych = 6,094%)

Ustaw. domyślnych param. MARSplines/MAPE [%] 2009-2014									
Ust.	5 1 2	10 1 2	15 1 2	20 1 2	25 1 2	30 1 2	35 1 2	40 1 2	
MAPE	6,210	6,139	6,110	6,107	6,092	6,090	6,099	6,099	
Ust.	30 2 2	30 3 2	30 4 2	30 5 2	30 6 2	-	-	-	
MAPE	6,008	5,989	5,999	5,997	5,997	-	-	-	
Ust.	30 3 3	30 3 4	30 3 5	-	-	-	-	-	
MAPE	6,016	6,047	6,047	-	-	-	-	-	

Legenda: Ust. – kolejne wartości oznaczają maksymalną liczbę funkcji bazowych, rząd interakcji oraz karę.

Tabela 6. Błędy MAPE *ex post* (wszystkie zm. obj.) bez usuwania nieistotnych FB (błąd MAPE dla ustawień domyślnych = 6,084%)

Ustaw. domyślnych param. MARSplines/MAPE [%] 2009-2014									
Ust.	5 1 2	10 1 2	15 1 2	20 1 2	25 1 2	30 1 2	35 1 2	40 1 2	
MAPE	6,207	6,137	6,107	6,097	6,086	6,088	6,084	6,084	
Ust.	35 2 2	35 3 2	35 4 2	35 5 2	35 6 2	-	-	-	
MAPE	5,966	5,964	5,948	5,953	5,953	-	-	-	
Ust.	35 4 3	35 4 4	-	-	-	-	-	-	
MAPE	5,948	5,948	-	-	-	-	-	-	

Legenda: tak jak dla tabeli 5.

Analiza tabel 5 i 6 wskazuje, że dezaktywacja opcji usuwania nieistotnych FB daje korzystniejsze wyniki. Wybór wszystkich zmiennych zamiast zmiennych typowanych przez analizowaną metodę w analizowanym przykładzie dawał większą dokładność prognozowania wygasłego.

Do przeprowadzenia prognoz *ex ante* wybrano model wykorzystujący wszystkie dostępne zmienne objaśniające z uwzględnieniem braku usuwania nieistotnych FB. Model ten posiada 35 funkcji bazowych, rząd interakcji wynoszący 4 oraz współczynnik kary wynoszący 2. Zbiorem testowym przeznaczonym do przeprowadzenia prognoz *ex ante* był szereg czasowy opisujący wszystkie zmienne w pierwszej godzinie doby 2015 roku. Błąd MAPE w trybie *ex ante* uzyskany dla tego roku przy zastosowaniu modelu bazującego na danych historycznych za lata 2009-2014 wyniósł 6,57% co jest wynikiem mniej korzystnym o 0,62% w porównaniu do badań uzyskanych w trybie *ex post*. W dalszych badaniach dodano zmienne jakościowe objaśniające (1, 2, 5, 7) oraz usunięto obserwacje odstające poprzez zastąpienie ich wartością z poprzedniej obserwacji (tab. 7 dla lat 2009-2014 i tab. 8 dla roku 2015).

Tabela 7. Błędy MAPE *ex post* (wszystkie zm. obj.) dla ustawień tab. 6 ze zmiennymi jakościowymi (ZJ), kodowaniem binarnym zmiennych wejściowych i usuwaniem obserwacji odstających.

Ustaw. param. MARSplines/MAPE [%] 2009-2014								
Ust.	D	DM	DM	D	DM	DM	DM	DM
	25-1-2 (**)	25-1- 2 (**)	35-4-2 (**)	25-1-2 (*)	25-1-2 (**)	25-1-2 (**)	35-4-2 (**)	35-4-2 (**)
MAPE	3,70	3,49	3,39	6,04	3,67	3,49	3,36	3,45

Legenda: * - usunięcie obserwacji odstających; ** - ZJ – wprowadzenie zmiennych jakościowych; D – dni binarne; DM – dni i miesiące binarne; *** - OO – usunięcie obserwacji odstających (zastąpienie wartością z poprzedniej obserwacji). Uwaga: Dane dla okresu 2009-2014.

Tabela 8. Błędy MAPE *ex post* (wszystkie zm. obj.) dla ustawień tab. 6 ze zmiennymi jakościowymi (ZJ), kodowaniem binarnym zmiennych wejściowych i usuwaniem obserwacji odstających.

Ustaw. param. MARSplines/MAPE [%] 2015								
Ust.	D	DM	DM	D	DM	DM	DM	DM
	25-1-2 (**)	25-1- 2 (**)	35-4-2 (**)	25-1-2 (*)	25-1-2 (**)	25-1-2 (**)	35-4-2 (**)	35-4-2 (**)
MAPE	2,62	2,55	2,13	5,54	2,63	2,55	2,13	2,29

Legenda: tak jak dla tabeli 7. Uwaga: Dane dla roku 2015.

Wnioski

Wykonane analizy potwierdziły wrażliwość metody na występowanie obserwacji odstających i dopasowanie do szeregu czasowego w trybie prognozowania wygasłego. Wykazano też, że znaczący wpływ na poprawę jakości prognoz *ex ante* ma wprowadzenie zmiennych

objaśniających jakościowych oraz kodowanie binarne zmiennych objaśniających ilościowych. Wybór najkorzystniejszych ustawień i ich weryfikacja w *ex ante* za 2015 rok dało błąd MAPE=5,03% przy wyłącznym uwzględnieniu wybranych predyktorów jakościowych oraz 4,93% przy dodatkowym zastąpieniu obserwacji odstających obserwacjami za poprzednią dobę. Analizy wykazały, że metoda MARSplines stanowi ciekawe narzędzie prognostyczne. Temperatura otoczenia ma największy wpływ na zapotrzebowanie na energię elektryczną w KSE. Opracowane prognozy wygasłe, z uwzględnieniem wrażliwych zachowań tej metody, dają obiecujący poziom dokładności. Prognozowanie *ex ante* dało wyniki ponad 2-krotnie mniej dokładne, co jest na obecnym etapie badań mniej korzystne niż skuteczność rzeczywistego prognozowania *ex ante* dla zapotrzebowania mocy w KSE dla analizowanej godziny doby. Dla modeli wyrównywania wykładniczego bez trendu, Holta, Wintersa (addytywnego i multiplikatywnego), uzyskano za 2015 r. *ex post* błędy MAPE odpowiednio o wartościach 4,63%, 6,83%, 8,64% i 8,66%, wykazując tym korzystne rezultaty metody MARSplines. W przyszłych badaniach niezbędne będzie wykorzystanie nowej grupy zmiennych objaśniających (po wykonaniu testu istotności zmiennych F Snedecora) oraz analiza pozostałych godzin doby.

Autorzy: mgr inż. Rafał Czapał, PSE Innowacje Sp. z o.o., Centrum Kompetencji Badania i Rozwój, ul. Jordana 25, 40-056 Katowice, E-mail: rafal.czapal@pse.pl; mgr inż. Pablo Benalcázar, Instytut Gospodarki Surowcami Mineralnymi i Energią Polskiej Akademii Nauk, Kraków, ul. J. Wybickiego 7A, 31-261 Kraków, E-mail: benalcazar@min-pan.krakow.pl; dr hab. inż. Jacek Kamiński, prof. IGSMiE PAN, Instytut Gospodarki Surowcami Mineralnymi i Energią Polskiej Akademii Nauk, Kraków, ul. J. Wybickiego 7A, 31-261 Kraków, E-mail: kaminski@min-pan.krakow.pl.

LITERATURA

- [1] Friedman J. H., Multivariate Adaptive Regression Splines, *The Annals of Statistics*, (1991), 19:1, 1-141
- [2] Friedman J. H., Estimating Functions of Mixed Ordinal and Categorical Variables Using Adaptive Splines, *New Directions in Statistical Data Analysis and Robustness* (Morgenthaler, Ronchetti, Stahel, eds.), Birkhauser, (1993)
- [3] Friedman J. H., Fast MARS, *Stanford University Department of Statistics, Technical Report 110*, (1993)
- [4] Witryna internetowa: www.statsoft.pl (dostęp: 2018.06.16)
- [5] Sokołowski A., Pasztyła A., Data Mining w prognozowaniu zapotrzebowania na nośniki energii, *StatSoft Polska Sp. z o. o.*, (2004), 91-102
- [6] Sokołowski A., Metody stosowane w Data Mining, *StatSoft Polska Sp. z o. o.*, (2002), 5-12
- [7] Wątroba G., Przykład rozwiązania zagadnienia predykcyjnego za pomocą technik Data Mining, *StatSoft Polska Sp. z o. o.*, (2002), 83-94
- [8] Migut G., Czy stosowanie metod Data Mining może przynosić korzyści w badaniach naukowych? *StatSoft Polska Sp. z o. o.*, (2009), 49-65
- [9] Nisbet N., Elder J., Miner G., *Handbook of Statistical Analysis & Data Mining*, Elsevier Inc., (2009), 83, 159-163
- [10] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, (2001), Springer
- [11] Söderström M., Sohlenius G., Rodhe L., Piikki K., Adaptation of regional digital soil mapping for precision agriculture, *Precision Agric/Springerlink*, (2016), 1-20
- [12] Lewicki P., Hill T., *Statistics: Methods and Applications*, 403-409
- [13] Witryna internetowa www.twojapogoda.pl (dostęp: 2016.01.14)
- [14] Witryna internetowa www.tvnmeteo.tvn24.pl (dostęp: 2016.01.14)
- [15] Witryna internetowa www.pse.pl (dostęp: 2016.01.14)