Rafał CZAPAJ[1], Pablo BENALCAZAR[2], Jacek KAMIŃSKI[2]

PSE Innowacje Sp. z o.o., Centrum Kompetencji Badania i Rozwój (1),
Instytut Gospodarki Surowcami Mineralnymi i Energią Polskiej Akademii Nauk, Kraków (2).

doi:10.15199/48.2019.08.03

Dobór zmiennych objaśniających z wykorzystaniem metody MARSplines na przykładzie prognozowania dobowego zapotrzebowania na moc szczytową 15-minutową w KSE

Streszczenie. *Artykuł prezentuje możliwość skorzystania z metod statystycznych automatyzujących dobór zmiennych objaśniających na przykładzie szczytowego obciążenia dobowego KSE. Testy ex post dotyczyły 10 zbiorów zmiennych objaśniających dla metod statystycznych klasycznych i typu Data Mining. Uzyskana macierz wyników pozwala wstępnie wybrać najkorzystniejszy zbiór zmiennych objaśniających i metodę statystyczną.*

Abstract*. The article examines the possibility of using statistical methods for the automated selection of explanatory variables of the daily peak demand in the National Power System. An analysis of 10 explanatory variable sets was conducted through classical and Data Mining methods. The obtained results, which are presented as a matrix of (ex-post) statistical measures, prove to be useful in the selection of the appropriate statistical method and the selection of explanatory variables. (**Explanatory Variable Selection Using the MARSplines Method on the Example of Forecasting the Daily 15-minute Peak Power Load in the National Power System**).*

Słowa kluczowe: MARSplines, zmienne objaśniające, prognozowanie, KSE, zapotrzebowanie na moc.
Keywords: MARSplines, Explanatory Variable, Forecasting, National Power System, Power Load.

Wstęp

Metoda MARSplines (ang. *Multivariate Adaptive Regression Splines*) stanowi wielozmienną adaptacyjną metodę regresyjną wykorzystującą funkcje sklejane (tzw. splajny). Autorem tej metody, która została zaproponowana w 1991 r., jest Jerome H. Friedman [1]. Metoda MARSplines oznaczona jest znakiem towarowym, a właścicielem jej licencji jest firma Salford Systems. Algorytm został zaprojektowany dla średnich zbiorów danych, które nie przekraczały 1000 obserwacji i dotyczyły maksymalnie 20 zmiennych objaśniających [1]. Przeprowadzone analizy wykonano na przykładzie prognozowania zapotrzebowania na moc szczytową w Krajowym Systemie Elektroenergetycznym (KSE). Moc szczytowa, w analizowanym przykładzie, stanowi najwyższą wartość 15-minutowego poboru mocy czynnej, z uwzględnieniem strat wynikających z jej przesyłu.

Zalety metody MARSplines

Zaletami metody MARSplines są [2, 3, 4]:

- rozwiązywanie problemów regresyjnych i klasyfikacyjnych;
- nieparametryczność natury metody skutkująca brakiem wymagania co do założeń nt. zależności pomiędzy zmiennymi niezależnymi i zmiennymi zależnymi (zależność liniowa, logistyczna itp.);
- elastyczność wyższa niż metod regresji liniowej;
- łatwość zrozumienia modeli i ich interpretacji;
- możliwość pracy z danymi ciągłymi i dyskretnymi;
- łatwiejsze operowanie na zmiennych numerycznych w wyniku braku stałej segmentacji danych;
- automatyczny dobór zmiennych;
- łatwość modelowania nieliniowości;
- łatwość modelowania interakcji pomiędzy zmiennymi co pozwala na lepsze odzwierciedlenie w prognozie czynników opisujących zjawisko (np. zróżnicowanie zapotrzebowania na moc w poszczególnych dniach tygodnia);
- łatwość pracy z dużymi zbiorami danych;
- łatwość pracy ze zmiennymi o dużym poziomie skomplikowania, niemonotoniczności oraz trudności w modelowaniu typu parametrycznego;
- szybkość budowania modeli dla dużych zbiorów danych (w szczególności w porównaniu do metody *Supported Vector Machines*, w której każda zmienna musi zostać przemnożona przez odpowiedni element lub wektor wspierający) dzięki wykorzystaniu szybkiej metody najmniejszych kwadratów.

Wady metody MARSplines

Wadami metody MARSplines są [2, 3]:

- konieczność usuwania danych odstających przed zastosowaniem tej metody;
- pewna doza arbitralności w doborze zmiennych (w szczególności w przypadku współliniowości tych zmiennych);
- zagrożenie nadmiernym dopasowaniem do danych (ponieważ metoda bazuje na metodzie rekurencyjnego podziału przestrzeni cech *recursive partitioning*);
- brak bezpośredniego wyliczania przedziałów ufności (i innych sprawdzeń modelu) w przeciwieństwie do modeli budowanych przy wykorzystaniu regresji liniowej, co powoduje konieczność walidacji krzyżowej modeli i powiązanych technik;
- dopasowanie do danych jest nieco gorsze niż w przypadku metody drzew wzmacnianych, jednakże istnieje znacząca łatwość w interpretacji wpływu każdej z analizowanych zmiennych.

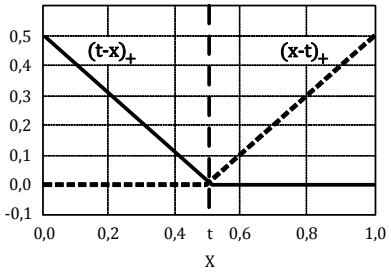
Podstawy metody MARSplines

Metoda MARSplines należy do grupy metod wykorzystujących *uczenie nieukierunkowane*. W przypadku takiego typu uczenia poszukiwany jest model najlepiej pasujący do danych bez uwzględnienia mechanizmu jaki te dane wygenerował. Poszukiwany model nie ma służyć opisowi relacji, związków przyczynowych, a tylko prognozowaniu lub rozpoznawaniu [5] (typowane zagadnienia *Data Mining* [6, 7, 8, 9, 10]). Omawiana metoda jest metodą ukierunkowaną (ang. *Supervised Learning*). Koncepcja metody rozszerza tradycyjne ujęcie zmiennych objaśniających w modelu regresyjnym. Poza całościowym uwzględnieniem wpływu predyktorów (tak jak w klasycznym modelu regresji) analizowane są wszystkie obserwacje danej zmiennej objaśniającej i obszar jej zmienności dzielony jest na przedziały, w których ma ona różny wpływ na badane zjawisko [5]. Do określania granic przedziałów stosowane są tzw. węzły (ang. *Knots*) stanowiące wartości progowe. Dzięki zastosowaniu takiego podejścia przeprowadzane jest porównanie wartości

zmiennej objaśniającej z wartością progową, czego efektem może być przyjęcie przez tę zmienną pewnej wartości wagi oraz różnego znaku. Wartość progowa określana jest jako *(t)*, a do rozróżnienia wartości zmiennej objaśniającej stosowana jest funkcja bazowa (ang. *Basis Function*) określana jako: $\beta imax(0; X-t)$ i została przedstawiona na rysunku 1 [5, 11, 12]. Ogólnie postać funkcji MARSplines (zależność 1) uzyskuje się poprzez sumowanie funkcji bazowych oraz iloczynów tych funkcji z odpowiednimi wagami, określonych wspólnie oznaczeniem $h_m(X)$:

$$y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m\left(X_{\upsilon(k,m)}\right) \quad (1)$$

gdzie: sumowanie przebiega przez wszystkie M składników modelu, a β_0 i β_m to parametry modelu (podobnie jak węzły t każdej funkcji bazowej, wyznaczane z danych) [4].



$$\text{gdzie: } (x-t)_+ = \begin{cases} x-t, dla\ x > t \\ 0, dla\ x \le t \end{cases}$$

Rys.1. Funkcja bazowa stosowana w metodzie MARSplines

Tak więc, $\underline{y}$ obliczane jest jako funkcja zmiennych predykcyjnych X (i ich interakcji) [4]. Elementami tej funkcji są, rzędna początkowa (β_0) i ważona (wagami β_m) suma jednej lub wielu funkcji bazowych $h_m(X)$, takich jak pokazane na rysunku 1 [4]. Na model ten można patrzeć jak na ważoną sumę funkcji bazowych, wybranych ze zbioru dużej liczby takich funkcji, pokrywających wszystkie wartości, każdego z predyktorów (w zbiorze tym mamy funkcję bazową i parametr t dla każdej, poszczególnej wartości, każdego predyktora) [10]. Algorytm MARSplines przeszukuje przestrzeń wszystkich wartości wejściowych i predykcyjnych (położeń węzłów *t*), jak i interakcji między zmiennymi [4]. Do modelu dodawane są wtedy kolejne funkcje bazowe (wybierane ze zbioru wszystkich dopuszczalnych funkcji) w taki sposób, by maksymalizować ogólny poziom dopasowania (wg minimum sumy kwadratów) [10]. Wynikiem tej operacji jest znalezienie najważniejszych zmiennych niezależnych, oraz najważniejszych ich interakcji [4].

Funkcja H w [4] określona jest formułą (2) następująco:

$$H_{km}\left(x_{\upsilon(k,m)}\right) = \prod_{k=1}^{K} h_{km} \quad (2)$$

gdzie: $x_{v(k,m)}$ jest predyktorem k w m-tym iloczynie.

Dla rzędu interakcji K=1 otrzymuje się model addytywny, dla rzędu K=2 model jest parami interakcyjny [4]. W krokowej procedurze postępującej, do modelu dodawane są kolejne funkcje bazowe, do zadanej, maksymalnej liczby, która powinna być dostatecznie duża (co najmniej dwa razy większa od optymalnej, pod względem minimum kwadratów) [4]. Model tworzony przy wykorzystaniu metody MARSplines w sensie teoretycznym stanowi sumę wyselekcjonowanych funkcji bazowych z dostępnego zestawu wszystkich wartości zmiennych objaśniających. Tym samym uzyskiwana liczba dopuszczalnych funkcji bazowych wynosi $2Nk$ (gdzie: N – liczba obserwacji, k – liczba predyktorów, 2 – konsekwencja zastosowania znaku + lub -). Algorytm metody dokonuje przeszukiwania przestrzeni obserwacji. Skutkiem przeszukiwania są wyznaczone wartości progowe (węzły) oraz wzajemne zależności pomiędzy zmiennymi objaśniającymi. Następnie wykonywana jest operacja budowania funkcji bazowych w oparciu o wyznaczone węzły i wagi. Zastosowanie podziału zbioru funkcji bazowych i wag, stanowiących podstawę tej metody, na podobszar funkcji klasyfikacyjnych i regresyjnych, określa jej nieparametryczny charakter. W wyniku powyższego, analiza zależności o złożonej i niemonotonicznej naturze może dawać korzystne rezultaty. Przeszukiwanie przestrzeni obserwacji działa w oparciu o metodę rekurencyjnego podziału przestrzeni cech (ang. *Recursive Partitioning*) [1] działającego naprzemiennie w dwóch etapach. W etapie pierwszym dodawane są funkcje bazowe w celu zwiększenia stopnia złożoności modelu. Zachodzi to do chwili wykorzystania wszystkich funkcji wprowadzonych na wejście budowanego modelu. W drugim etapie (opcjonalnym) uruchamiana jest funkcja przycinania (ang. *Prunning*). Przycinanie polega na usuwaniu najmniej istotnych funkcji bazowych z modelu. Usunięciu podlegają tylko te funkcje bazowe, które nie spowodują znaczącej utraty dobroci dopasowania (w sensie najmniejszych kwadratów). Etap drugi zostanie zakończony po osiągnięciu minimalnej wartości współczynnika GCV (ang. *Generalized Cross Validation Error*), który stanowi uogólniony błąd stosowany w ocenie krzyżowej. Zaletą stosowania współczynnika GCV (zależność (3)), będącego miarą dobroci dopasowania modelu do danych rzeczywistych, jest uwzględnienie nie tylko wielkości błędu resztowego, ale również stopnia złożoności modelu [5]:

$$GCV = \frac{\sum_{i=1}^{N}\left(y_i - f(x_i)\right)^2}{\left(1-\frac{C}{N}\right)^2}; \text{ gdzie: } C=1+cd \quad (3)$$

gdzie: N – liczba przypadków w zbiorze danych, d – liczba stopni swobody (równa liczbie niezależnych funkcji bazowych), c – kara za dodanie do modelu kolejnej funkcji bazowej. Z doświadczenia wynika, że najlepsze C otrzymuje się przy 2<d<3 [4, 10].

Zbiór zmiennych objaśniających – zmienne meteorologiczne

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 (poniżej ±1°C) z obu biegunów temperatury [13, 14]. Pierwszy zestaw zmiennych (najliczniejszy) składa się z 14 następujących parametrów meteorologicznych:

- temperatura otoczenia maksymalna (°C) - (Z6);
- temperatura otoczenia minimalna (°C) - (Z7);
- opady deszczu (mm) - (Z8);
- prędkość wiatru średnia (km/h) - (Z9);
- prędkość wiatru średnia (km/h) ekspercka - (Z10);
- prędkość wiatru maksymalna (km/h) - (Z11);
- ciśnienie atmosferyczne (hPa) - (Z12);
- liczba stopniodni grzewczych (°Cdzień) - (Z13);
- liczba stopniodni chłodniczych (°Cdzień) - (Z14);
- liczba godzin słonecznych (-) - (Z15);
- energia słoneczna (W/m^2) - (Z16);
- poziom promieniowania UV (-) - (Z17);
- temperatura punktu Rosy (°C) - (Z18);
- temperatura mokrego termometru (°C) - (Z19).

Do powyższego zestawu dołączono dwie zmienne objaśniające, które istotnie uzupełniają powyższy zbiór. Pierwsza z nich 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 ma postać czasową niezakodowaną, w której zawarto informację o czasie (z rozdzielczością 15-minutową) wystąpienia szczytowej wartości mocy 15-minutowej obciążenia KSE - dane te pozyskano z publikacji PSE S.A. [15] w każdej dobie analizowanego szeregu czasowego. Zmienną objaśnianą była wspomniana szczytowa wartość mocy 15-minutowej dobowego obciążenia KSE.

Zbiór zmiennych objaśniających – pozostałe zmienne

Dodatkowo, w celu uwydatnienia wpływu warunków wiatrowych na obciążenie KSE, przetestowano zmienne Z20-23, a także podjęto próbę oceny przydatności zakodowanej informacji o cyklu księżyca do wyjaśniania zmiennej objaśnianej. Do zmiennych tych zaliczono poniższe pozycje:

- udział procentowy dobowej szczytowej mocy 15-minutowej w szczycie tygodniowym - (Z4);
- zakodowana informacja o fazie księżyca - (Z5);
- maksymalna generacja w farmach wiatrowych - (Z20);
- godzina wystąpienia maksymalnej generacji w farmach wiatrowych - (Z21);
- moc osiągalna w farmach wiatrowych - (Z22);
- procentowy udział mocy generowanej w farmach wiatrowych do ich mocy zainstalowanej - (Z23).

Wytypowany zbiór zmiennych objaśniających stanowi Główny Zbiór Zmiennych Objaśniających (GZZO). Poszczególne zmienne stanowią ciąg danych historycznych charakteryzujących się rozdzielczością dobową. Obejmują okres 5 lat, od 1 stycznia 2010 r. do 31 grudnia 2014 r.

Wytypowane alternatywne metody statystyczne

Jako alternatywne metody statystyczne (S) dla metody MARSplines (S1), w celach porównawczych, wytypowano „klasyczne" (na podstawie klasyfikacji zgodnej z układem w środowisku Statistica) metody doboru zmiennych objaśniających, do których należą:

- Wielozmienna regresja adaptacyjna z użyciem funkcji sklejanych (*Multivariate Adaptive Regression Splines*) - (S1).
- Regresja wieloraka (*Multiple Regression*) - (S2).
- Ogólne modele liniowe i nieliniowe (*Generalized Linear and Non-Linear Models*) - (S3).
- Ogólne modele regresji (*General Regression Models*) - (S4).
- Modele cząstkowych najmniejszych kwadratów (*Fragmentary Least Squares Models*) - (S5).
- Sztuczne Sieci Neuronowe (*Artificial Neural Networks*) - (S6).

Powyższe metody były testowane w trybie *ex post* (prognozy wygasłe na znanych wartościach historycznych) pod kątem jakości dopasowania zmiennych objaśniających do zmiennej objaśnianej. W tym celu przy pomocy metod określono 10 zestawów zmiennych objaśniających wytypowanych z GZZO automatycznie (przez niektóre z poniższych metod) i ręcznie metodą eksperckią.

Do metod (M_n) zautomatyzowanych zaliczono (z odpowiednim kryterium):

- regresję wieloraką ($\beta>\pm0,04$) - (M1);
- metodę MARSplines (ranking predyktorów) - (M2);
- obliczanie współczynnika Pearsona (>0,47) - (M4);
- metodę szybką C&RT (ranking predyktorów) - (M8);
- dobór i eliminację zmiennych (ranga zmiennej) - (M10).

Do metod eksperckich zaliczono (M_n):

- wybór zmiennych objaśniających postrzeganych jako posiadające znaczący wpływ na obciążenie KSE - (M3);
- wybór wszystkich posiadanych zmiennych objaśniających (brak kryterium) - (M5);
- wybór wszystkich spośród najlepszych zmiennych objaśniających z metod M1-M4 (zgodnie z kryteriami dla metod od M1 do M4) - (M6);
- wybór wszystkich spośród najlepszych zmiennych objaśniających oraz ekspercki dobór dodatkowej/dodatkowych zmiennych objaśniających - (M7);
- regresję wieloraką przeprowadzaną iteracyjnie osobno dla każdej zmiennej objaśnianej ($\beta>\pm0,1$) - (M9).

Uzyskane zestawy zmiennych objaśniających

Po zastosowaniu opisanego powyżej podejścia uzyskano następujące zestawy (M) zmiennych objaśniających (Z):

- Z (4, 6-7, 13, 16-17, 19) - 7 zmiennych - M1;
- Z (3-4, 6-7, 16-17, 20, 22) - 8 zmiennych - M2;
- Z (11, 13-15, 18) - 5 zmiennych - M3;
- Z (4, 6-7, 13, 16-17) - 6 zmiennych - M4;
- Z (1, 3-23) - 22 zmienne - M5;
- Z (3-4, 6-7, 11, 13-20, 22) - 14 zmiennych - M6;
- Z (3-4, 6-7, 11-20, 22) /dodatkową zmienną jest ciśnienie - Z12/ - 15 zmiennych - M7;
- Z (4, 6-7, 12-20, 22) - 13 zmiennych - M8;
- Z (4, 6-7, 13-14, 16-19, 22-23) - 11 zmiennych - M9;
- Z (3-4, 6-7, 13-14, 16-19) - 10 zmiennych - M10.

Zastosowane mierniki *ex post*

Wytypowano 1 metodę statystyczną należącą do grupy metod *Data Mining* oraz 5 klasycznych metod prognostycznych i następnie poddano testom *ex post*. W testach wykorzystano 10 zbiorów zmiennych objaśniających należących do zbioru GZZO. Rezultatem testów jest macierz wyników. Każdy z elementów tej macierzy stanowi średnią arytmetyczną czterech procentowych mierników *ex post* jakości dopasowania danych do modelu, tj. [16, 17, 18, 19, 20]:

- MPE - Średni błąd procentowy (*Mean Percentage Error*).
- MAPE - Średni całkowity błąd procentowy (*Mean Absolute Percentage Error*).
- RMSPE - Pierwiastek średniokwadratowego błędu procentowego (*Root Mean Square Percentage Error*).
- Theil - Współczynnik rozbieżności Theila.

Poziom wartości mierników *ex post*, który uznano za kwalifikujący do dalszych rozważań, ustalono na poziomie nieprzekraczającym 5%. Założono również, że za najkorzystniejsze uznane zostaną zestawy zmiennych objaśniających gwarantujące jak najniższą średnią arytmetyczną ze wspomnianych wartości mierników *ex post* dla każdego roku z osobna. Analogiczne założenie przyjęto dla wytypowania najkorzystniejszej metody/metod, które będą uznane za najlepiej rokujące dla przyszłych badań, w oparciu o zakres danych uwzględnionych w tej publikacji.

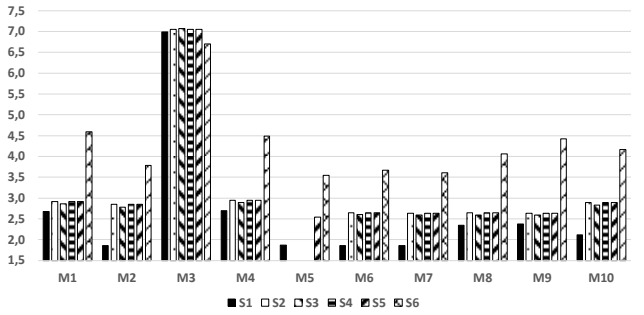
Analiza uzyskanych wyników uśrednionych

Macierz wyników testów *ex post* wstępnie wytypowanych metod statystycznych i wstępnie wytypowanych zestawów zmiennych objaśniających zaprezentowano w tabeli 1. W wierszach zamieszczono wyniki uzyskiwane dla wybranych metod statystycznych oferowanych w pakiecie *Statistica*, a w kolumnach zamieszczono wyniki uzyskane dla poszczególnych zestawów zmiennych objaśniających. Odzwierciedleniem graficznym tabeli 1 jest rysunek 1.

Metodą doboru zmiennych objaśniających, która średnio najlepiej określiła zestaw takich zmiennych, jest metoda (M5). Uzyskana średnia arytmetyczna wszystkich uśrednionych mierników (dla każdego roku z osobna) jakości szacunków *ex post* wyniosła 2,65%. Kolejne dwa miejsca zajęły metody M7 oraz M6 ze średnim wynikiem na poziomie odpowiednio 2,66% oraz 2,68%. Kolejne trzy miejsca zajęły odpowiednio metody M8, M2 oraz M9, które uzyskały średnie wyniki na poziomie odpowiednio: 2,82%; 2,83% oraz 2,88%. Średnie wyników na poziomie 2,97%; 3,14% oraz 3,15% uzyskane zostały poprzez wykorzystanie metod odpowiednio M10; M1 oraz M4. Ostanie, dziesiąte, miejsce zajęła w analizowanym przypadku metoda M3 ze średnim wynikiem na poziomie 6,99%.

Tabela 1. Macierz uśrednionych mierników procentowych *ex post* z każdego roku z osobna - metoda statystyczna x metoda doboru zmiennych objaśniających (S x M) (gdzie: Śr - średnia)

S\M	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	Śr.
S1	2,68	1,86	6,99	2,70	1,87	1,86	1,86	2,34	2,38	2,12	2,67
S2	2,91	2,85	7,05	2,94	-	2,65	2,63	2,64	2,63	2,89	3,24
S3	2,86	2,78	7,07	2,89	-	2,60	2,59	2,59	2,59	2,83	3,20
S4	2,91	2,85	7,05	2,94	-	2,65	2,63	2,64	2,63	2,89	3,24
S5	2,91	2,85	7,05	2,94	2,54	2,65	2,63	2,64	2,63	2,89	3,17
S6	4,59	3,78	6,70	4,49	3,54	3,67	3,61	4,06	4,42	4,17	4,30
Śr.	3,14	2,83	6,99	3,15	2,65	2,68	2,66	2,82	2,88	2,97	3,28



Rys.1. Uśrednione wartości mierników *ex post* na tle analizowanych metod statystycznych

Metodą statystyczną, która średnio najlepiej operowała na typowanych zestawach zmiennych objaśniających, była metoda S1 (MARSplines) z wynikiem 2,67%. Drugie i trzecie miejsce w tej klasyfikacji zajęły metody S5 (3,17%) oraz S3 (3,20%). Najsłabiej na typowanych zestawach operowała metoda wykorzystująca sztuczne sieci neuronowe S6 (4,30%). Najniższe cząstkowe wyniki uzyskiwane były przy „skrzyżowaniu" metod statystycznej S1 (MARSplines) z typowaniami oferowanymi przy podejściu stosowanym dla metod doboru zmiennych M2 (MARSplines), M6 oraz M7 (z na poziomie odpowiednio 1,86%) oraz M5 (1,87%). Wyniki te znacząco wyróżniają się na tle pozostałych par metod S oraz metod M.

Wnioski

Podstawowym celem prowadzonych badań, których wybrane wyniki zostały przedstawione w niniejszym artykule, było zautomatyzowanie i wytypowanie optymalnego zestawu zmiennych objaśniających i optymalnej metody statystycznej. Mając na uwadze powyższe zaleca się wybrać metodę M2 (MARSplines) typowania zestawu zmiennych oraz metodę statystyczną S1 (również MARSplines).

Należy zauważyć ciekawą zależność, że wykonane badania i analiza porównawcza wyników przeprowadzona na podstawie macierzy zamieszczonej w tabeli 1 wskazują, że najkorzystniej jest skupić się na metodzie MARSplines. 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 pamiętać jednakże na obecnym etapie, że metoda MARSplines na etapie uczenia (prognozowanie w trybie *ex post*) jest podatna na przeuczenie i w zderzeniu z realnym prognozowaniem może dawać mniej obiecujące wyniki. Autorzy pragną podkreślić, że kolejne badania prowadzone w tym zakresie mogą obejmować również dalsze poszukiwanie zmiennych objaśniających (nie tylko meteorologicznych), optymalizację długości ciągu czasowego umożliwiającego uzyskanie wyników na podobnym poziomie oraz testowanie wybranych metod w trybie *ex ante*.

***Autorzy**: mgr inż. Rafał Czapaj, PSE Innowacje Sp. z o.o., Centrum Kompetencji Badania i Rozwój, ul. Jordana 25, 40-056 Katowice, E-mail: rafal.czapaj@pse.pl; mgr inż. Pablo Benalcazar, 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
[14] Witryna internetowa www.tvnmeteo.tvn24.pl
[15] Witryna internetowa www.pse.pl
[16] Zeliaś A., Pawełek B., Wanat S., Prognozowanie ekonomiczne-Teoria, przykłady, zadania, *Wydawnictwo Naukowe PWN,* Warszawa (2003)
[17] Witkowska D., Podstawy ekonometrii i teorii prognozowania (Podręcznik z przykładami i zadaniami), *Oficyna Ekonomiczna*, Kraków (2005)
[18] Kot S.M., Jakubowski J., Sokołowski A., Statystyka - Podręcznik dla studiów ekonomicznych, *Centrum Doradztwa i Informacji,* Warszawa (2007)
[19] StatSoft Polska, STATISTICA - Przewodnik, *Wydawnictwo StatSoft Polska Sp. Z o.o.*, Kraków (2008)
[20] Helt P., Parol M., Piotrowski P., Metody sztucznej inteligencji w elektroenergetyce, *Oficyna Wydawnicza Politechniki Warszawskiej*, Warszawa (2000)