NAFTA-GAZ

Nafta-Gaz 2019, nr 2, s. 111–117, DOI: 10.18668/NG.2019.02.07

Prognozowanie zapotrzebowania na gaz metodami sztucznej inteligencji

Forecasting gas demand using artificial intelligence methods

Andrzej Paliński

AGH Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie, Wydział Zarządzania, Katedra Informatyki Stosowanej

STRESZCZENIE: W artykule zaprezentowano współczesne tendencje dotyczące metod sztucznej inteligencji i uczenia maszynowego, do których zalicza się między innymi sztuczne sieci neuronowe, drzewa decyzyjne, systemy oparte na logice rozmytej i inne. Metody inteligencji obliczeniowej stanowią część obszaru badań nad sztuczną inteligencją. Wybrane metody inteligencji obliczeniowej zostały wykorzystane do budowy średnioterminowych miesięcznych prognoz zapotrzebowania na gaz dla Polski. Porównana została trafność prognoz uzyskanych za pomocą sztucznej sieci neuronowej i drzewa decyzyjnego z klasyczną regresją liniową z wykorzystaniem danych historycznych z okresu dziesięciu lat. Jako zmienne objaśniające zastosowano: zużycie gazu w innych krajach UE, średnią miesięczną temperaturę, produkcję przemysłową, wynagrodzenie w gospodarce i cenę gazu. Prognozowanie przeprowadzono w pięciu etapach różniących się doborem próby uczącej i testowej, zastosowaniem wstępnego przygotowania danych oraz eliminacją niektórych zmiennych. W przypadku danych nieprzygotowanych i losowego zbioru uczącego najwyższą trafność osiągnęła regresja liniowa. W przypadku danych uzupełnionych i losowego zbioru uczącego najwyższą trafność uzyskano za pomocą drzewa decyzyjnego. Prognoza zbudowana na podstawie pierwszych ośmiu lat i testowana na dwóch ostatnich została najtrafniej utworzona za pomocą regresji, ale tylko nieznacznie lepiej niż przy wykorzystaniu drzewa decyzyjnego lub sieci neuronowej, niezależnie od normalizacji danych i eliminacji współliniowych zmiennych. Metody uczenia maszynowego wykazały się dobrą trafnością prognoz miesięcznego zużycia gazu, niemniej jednak nieznacznie ustąpiły klasycznej regresji liniowej ze względu na zbyt wąski zbiór zmiennych objaśniających. Metody uczenia maszynowego będą mogły wykazać się wyższą skutecznością wraz ze wzrostem liczby danych oraz rozszerzeniem zbioru potencjalnych zmiennych objaśniających. W morzu danych metody uczenia maszynowego są w stanie skuteczniej tworzyć modele prognostyczne bez konieczności żmudnej ingerencji analityka w przygotowanie danych i wieloetapową analizę. Pozwolą także na dowolnie częstą aktualizację postaci modeli prognostycznych, nawet po każdym uzupełnieniu zbioru danych.

Słowa kluczowe: prognozowanie, sztuczna inteligencja, zapotrzebowanie, gaz ziemny.

ABSTRACT: The paper presents contemporary trends in artificial intelligence and machine learning methods, which include, among others, artificial neural networks, decision trees, fuzzy logic systems and others. Computational intelligence methods are part of the field of research on artificial intelligence. Selected methods of computational intelligence were used to build medium-term monthly forecasts of natural gas demand for Poland. The accuracy of forecasts obtained using the artificial neural network and the decision tree with classical linear regression was compared based on historical data from a ten-year period. The explanatory variables were: gas consumption in other EU countries, average monthly temperature, industrial production, wages in the economy and the price of natural gas. Forecasting was carried out in five stages differing in the selection of the learning and testing sample, the use of data preprocessing and the elimination of some variables. For raw data and a random training set, the highest accuracy was achieved by linear regression. For the preprocessed data and the random learning set, the decision tree was the most accurate. The forecast obtained on the basis of the first eight years and tested on the last two was most accurately created by regression, but only slightly better than with the decision tree or neural network, regardless of data normalization and elimination of collinear variables. Machine learning methods showed good accuracy of monthly gas consumption forecasts, but nevertheless slightly gave way to classical linear regression, due to too narrow set of explanatory variables. Machine learning methods will be able to show higher effectiveness as the number of data increases and the set of potential explanatory variables is expanded. In the sea of data, machine learning methods are able to create prognostic models more effectively, without the analyst's laborious involvement in data preparation and multi-stage analysis. They will also allow for the frequent updating of the form of prognostic models even after each addition of new data into the database.

Key words: forecasting, artificial intelligence, demand, natural gas.

Autor do korespondencji: A. Paliński, e-mail: *palinski@zarz.agh.edu.pl*

Artykuł nadesłano do Redakcji 14.11.2018 r. Zatwierdzono do druku 11.02.2019 r.

Wstęp

Badania nad sztuczną inteligencją rozpoczęto już w latach 50. ubiegłego wieku, ale XXI wiek przyniósł gwałtowne ich przyspieszenie ze względu na wzrost mocy obliczeniowej komputerów, rozwój Internetu i lawinowy przyrost danych w bazach danych na całym świecie. Sztuczna inteligencja to dział informatyki zajmujący się tworzeniem modeli inteligentnych zachowań, które mogą być implementowane w programach komputerowych. Problemy rozwiązywane przy użyciu sztucznej inteligencji są problemami, których nie jesteśmy w stanie efektywnie rozwiązywać za pomocą klasycznych algorytmów. Sztuczna inteligencja łączy wiedzę z różnych obszarów, takich jak: informatyka, logika, robotyka, neurobiologia, psychologia i inne.

Inteligencja obliczeniowa (inaczej obliczenia inteligentne albo metody inteligencji obliczeniowej) jest częścią dziedziny sztucznej inteligencji. Inteligencja obliczeniowa to grupa algorytmów heurystycznych, takich jak: systemy oparte na logice rozmytej, sztuczne sieci neuronowe i obliczenia ewolucyjne. Do metod inteligencji obliczeniowej zaliczyć można także uczenie maszynowe, metody statystyczne, modelowanie bayesowskie, teorię zbiorów przybliżonych i metody drążenia danych (ang. *data mining*). Cechą wspólną większości modeli inteligencji obliczeniowej jest to, że są to metody uczenia na podstawie danych. Algorytmy modyfikują swoje parametry, korzystając z napływających danych. Inteligencję obliczeniową nazywa się także obliczeniową sztuczną inteligencją lub określa terminem *soft-computing*.

Jednym z obszarów inteligencji obliczeniowej jest uczenie maszynowe, którego celem jest automatyzacja procesu budowy rozwiązania problemu. Istnieje wiele algorytmów uczenia maszynowego i różne kryteria ich podziału. Jednym z nich jest podział oparty na sposobie uczenia: uczenie nadzorowane i nienadzorowane. Uczenie nadzorowane odbywa się na jasno opisanym zbiorze danych zawierającym poprawne wyniki wnioskowania, najczęściej przy problemach klasyfikacji. W przypadku uczenia nienadzorowanego wynik końcowy nie jest znany, działanie algorytmu polega na szukaniu niedostrzegalnych zależności w wielowymiarowych zbiorach danych. Przykładem jest grupowanie i organizacja. Drugi najczęściej spotykany podział oparty jest na typie problemu rozwiązywanego przez dany algorytm. W tym przypadku wyróżnia się: algorytmy regresyjne, klasyfikacyjne i grupujące. Dobór odpowiedniego algorytmu uzależniony jest nie tylko od rozpatrywanego problemu, ale również od ilości danych oraz liczby wymiarów je opisujących, a także od czasu uczenia, dokładności i liniowości. Najczęściej stosowane algorytmy to (Larose, 2006):

- metoda *k*-średnich – algorytm grupujący, służący do analizy skupień. Umożliwia znalezienie i wyodrębnienie z nieuporządkowanego zbioru danych grup elementów najbardziej do siebie podobnych (tzw. skupień lub klastrów);
- algorytm *k*-najbliższych sąsiadów – metoda klasyfikacji i filtrowania danych, mogąca służyć między innymi do odpowiedzi na pytanie, do jakiej klasy (grupy) należy zaliczyć nowy element w zbiorze danych;
- drzewa decyzyjne (klasyfikacyjne) – algorytmy stosowane głównie w procesach podejmowania decyzji, w problemach klasyfikacji danych oraz predykcji. Każde drzewo składa się z korzenia, z którego prowadzą co najmniej dwie ścieżki do kolejnych węzłów. W każdym węźle, za pomocą instrukcji warunkowej, podejmowana jest decyzja o podziale na podzbiory zawierające instancje o podobnych wartościach. Algorytm kończy działanie, gdy dochodzi do węzła (liścia), z którego nie wychodzą już kolejne gałęzie;
- sztuczne sieci neuronowe – algorytmy wykorzystywane w większości zagadnień klasyfikacji, predykcji i grupowania. Model sieci neuronowej składa się z połączonych ze sobą neuronów, w których wagi połączeń z kolejnymi neuronami są ustalane w trakcie procesu uczenia sieci.

Algorytm inteligencji obliczeniowej i uczenia maszynowego, jak każdy model matematyczny, może zostać napisany w dowolnym języku programowania lub wykonany za pomocą programów MS Excel lub Mathcad. Do złożonych problemów i podczas programowania systemów uczących się można zastosować gotowe biblioteki programistyczne, oferujące zoptymalizowane i sprawdzone wersje algorytmów. Najpopularniejsze biblioteki napisane są w języku Python, i to właśnie ten język jest najczęściej kojarzony z programowaniem sztucznej inteligencji i analizą tekstów. Dużą popularność zdobył także język R, który stosowany jest głównie w odniesieniu do danych liczbowych.

Zaawansowaną analizę danych można przeprowadzić np. przy użyciu dodatków do MS Excel (Power Query oraz Data Mining Add-ins) albo rozbudowanego środowiska Microsoft Visual Studio z usługami Business Intelligence. Komercyjne oprogramowanie oferują również takie firmy jak: SAS, StatSoft (Statistica), SAP, KNIME, RapidMiner czy też IBM (SPSS). Niekiedy wygodniejsze i tańsze może być skorzystanie z gotowych komercyjnych narzędzi i zewnętrznych serwerów obliczeniowych (tzw. chmur). Dwie najpopularniejsze to Microsoft Azure Machine Learning i Amazon Machine Learning.

Celem artykułu jest ocena przydatności metod uczenia maszynowego w prognozowaniu zapotrzebowania na gaz w horyzoncie średnio- i długoterminowym. Dalszy układ artykułu jest następujący. W części drugiej przedstawiono charakterystykę stosowanych obecnie metod prognozowania zapotrzebowania

na gaz. W części trzeciej zaprezentowano przykłady budowy prognoz miesięcznego zapotrzebowania na gaz dla Polski z użyciem trzech algorytmów uczenia maszynowego. Praca zakończona jest krótkim podsumowaniem.

Prognozowanie zapotrzebowania na gaz

W literaturze przedmiotu najpopularniejszą grupą modeli używanych w prognozowaniu zapotrzebowania na gaz są modele statystyczno-ekonometryczne (np. Sánchez-Úbeda i Berzosa, 2007; Bianco et al., 2014a, 2014b). Klasyczne podejście do prognozowania zapotrzebowania na gaz wykorzystuje modele regresji i z góry określony zbiór danych wejściowych. Typowe zmienne objaśniające wykorzystywane w tradycyjnym prognozowaniu zapotrzebowania na gaz to (Balfe i Kelp, 2014):

- historyczne ceny gazu (lokalne, na rynkach światowych, bieżące i terminowe);
- dane pogodowe (temperatura, wiatr, opad i nasłonecznienie);
- liczba i zużycie gazu klientów indywidualnych (także współczynniki elastyczności cenowej i dochodowej);
- dane dotyczące budownictwa i działalności remontowej (budowa nowych domów, zmiany energochłonności i inne);
- liczba i zużycie gazu dużych odbiorców (wielkość i charakterystyka odbioru, współczynniki elastyczności);
- dane makroekonomiczne (PKB, inflacja, wynagrodzenie, stopa bezrobocia i inne);
- sytuacja polityczna;
- metadane (dane samorządu terytorialnego, urzędów statystycznych i inne).

Współczesne podejście do budowy modeli empirycznych wykracza poza tradycyjne modele statystyczno-ekonometryczne. Klasyczne modele zmuszają z jednej strony do spełnienia ścisłych wymagań dotyczących wnioskowania statystycznego, z drugiej strony – narzucają konieczność wyznaczenia z góry zbioru ewentualnych zmiennych objaśniających. Zbiór ten może być zawężony w toku wnioskowania, ale trudno go później rozszerzać.

Prognozowanie cen i zapotrzebowania na gaz może obecnie dzięki sztucznej inteligencji opierać się na dowolnie szerokim zbiorze potencjalnych zmiennych objaśniających, niekoniecznie związanych bezpośrednio z cenami gazu. Mogą to być dane uwzględniające częstość zapytań w wyszukiwarkach zawierających słowa kluczowe dotyczące gazu, rynku gazu, urządzeń gazowych itp., ale także: gazociągi, Rosja, dyrektywa gazowa itp. Jest to tzw. analiza *search volume* udostępniana przez Google. Istnieje również możliwość analizy pozytywnego lub negatywnego kontekstu wypowiedzi użytkowników portali społecznościowych i branżowych dotyczących gazu jako nośnika energii, ekologicznych źródeł energii, innych źródeł energii itp. Jest to tzw. analiza sentymentalna, która wykorzystuje słowniki zawierające informację o stopniu pozytywnego wydźwięku słów. Słowniki takie istnieją już także w języku polskim. Przydatne może być także grupowanie (klastrowanie) klientów według wzorców zachowań zakupowych (Gaweł et al., 2016).

Możliwość budowy modeli zawierających bardzo obszerny zbiór danych i wnioskowania wynikającego z danych zapewniają współczesne modele oparte na danych (ang. *data-driven models*), które wykorzystują wyniki badań z obszarów (Holdaway, 2014):

- sztucznej inteligencji;
- inteligencji obliczeniowej;
- *soft computing* – wnioskowanie oparte na pojęciach nieostrych – rozmyte systemy regułowe;
- uczenia maszynowego;
- eksploracji danych i odkrywania wiedzy w bazach danych.

Zaletą tego podejścia jest nie tylko pozostawienie doboru zmiennych algorytmom wbudowanym w oprogramowanie, ale także możliwość częstej aktualizacji postaci modelu (np. kwartalnej, miesięcznej lub częstszej). Dowolnie obszerny zakres danych może być gromadzony i stale aktualizowany w tematycznych hurtowniach danych, które będą ułatwiały częstą modyfikację modeli prognostycznych o nowe dane i nowe zmienne objaśniające (Paliński, 2018).

Ostatnie lata przyniosły duże zainteresowanie metodami eksploracji danych i sztucznej inteligencji (Suykens et al., 1996; Šebalj et al., 2017). Šebalj et al. (2017) w swoich badaniach dokonali porównania trafności ponad 400 modeli planowania zapotrzebowania na energię. Okazuje się, że metody statystyczne lepiej sprawdzają się w modelach krótko- i średniookresowych, natomiast w prognozowaniu długookresowym lepsze są metody oparte na inteligencji obliczeniowej. Wynika to m.in. z przewagi metod inteligencji obliczeniowej w tworzeniu modeli dla danych słabo oczyszczonych.

Prognoza miesięcznego zużycia gazu dla Polski

W dalszej części przedstawiony jest prosty przykład budowy prognozy zużycia gazu z wykorzystaniem typowego narzędzia do gromadzenia dużych zbiorów danych i wnioskowania na podstawie danych z użyciem narzędzi inteligencji obliczeniowej. Jest nim Microsoft SQL Server 2017 (Microsoft, b.d.-b) w wersji Enterprise, posiadającej rozbudowany moduł Business Intelligence i zaawansowanej eksploracji danych (Microsoft, b.d.-a).

Przy budowie modeli prognostycznych wykorzystano ogólnodostępne dane. Dane dotyczące miesięcznego zużycia gazu w krajach UE, w tym Polski, pochodzą z bazy Eurostat Energy

(Eurostat, b.d.). Dane te obejmują okres od stycznia 2008 r. do lutego 2018 r., co daje 122 obserwacje. Dane dotyczące średniej miesięcznej temperatury dla Warszawy, jako w przybliżeniu centralnego punktu Polski, wyliczono na podstawie dobowych danych pozyskanych z portalu FreeMeteo (FreeMeteo, b.d.). Ceny gazu ziemnego pochodzą ze strony internetowej IndexMundi (IndexMundi, b.d.). Co prawda nie dotyczą one bezpośrednio Polski, gdyż są cenami Henry Hub w Luizjanie, ale mogą stanowić punkt odniesienia dla światowych cen gazu, także cen w kraju oraz cen LNG. Ceny zakupu gazu przez podmioty zajmujące się obrotem gazem ziemnym w Polsce nie są powszechnie znane, ale w przypadku budowy systemu prognostycznego przez taki podmiot będzie znany jego własny koszt pozyskania gazu. Ponadto użyto danych GUS dotyczących średniego miesięcznego wynagrodzenia w gospodarce (Bank Danych Makroekonomicznych, b.d.) oraz miesięcznej produkcji przemysłowej (Bank Danych Lokalnych, b.d.). Wszystkie dane zapisano w bazie danych, która może być automatycznie aktualizowana.

Prognozowanie zostało wykonane w pięciu etapach. W pierwszym wykorzystano surowe dane bez ich wstępnego przetworzenia, które w przypadku posiadanych danych wymagałoby przede wszystkim uzupełnienia brakujących danych płacowych za lata 2008–2009 oraz nielicznych danych dotyczących zużycia gazu w niektórych krajach. W drugim etapie uzupełniono brakujące dane prognozami uzyskanymi metodą Wintersa i średniej ruchomej. W trzecim etapie zmieniono sposób oceny trafności prognoz, tworząc zbiór testowy na podstawie danych z przyszłych okresów. W czwartym etapie dodatkowo przeprowadzono normalizację zmiennych, a w piątym usunięto współliniowe zmienne.

W każdym etapie zbudowano trzy prognozy miesięcznego zużycia gazu. Każda z nich uzyskana została za pomocą innego algorytmu, kolejno: sztucznej sieci neuronowej, drzewa decyzyjnego i regresji liniowej.

We wszystkich etapach zmiennymi objaśniającymi były wstępnie:

- zużycie gazu w krajach UE charakteryzujących się najwyższym zużyciem: Niemcy, Francja, Hiszpania, Włochy, Holandia, Norwegia, Turcja, Wielka Brytania (jako element prognozowania analogowego);
- cena gazu Henry Hub;
- wskaźnik miesięcznej produkcji przemysłowej w cenach roku 2015;
- średnie miesięczne wynagrodzenie w gospodarce;
- średnia miesięczna temperatura dla Warszawy;
- rok i miesiąc.

Zużycie gazu w Polsce okazuje się silnie skorelowane ze zużyciem gazu w innych krajach europejskich – współczynniki korelacji liniowej powyżej 0,5, z wyjątkiem Norwegii (współczynnik korelacji nieistotnie statystycznie różny od zera). Podobna sytuacja do przypadku Norwegii dotyczy współczynnika korelacji zużycia gazu w Polsce z ceną gazu – współczynnik korelacji nieistotnie statystycznie różny od zera. Ponadto zużycie gazu w Polsce jest umiarkowanie skorelowane z wynagrodzeniem w gospodarce (współczynnik korelacji 0,3) oraz wartością produkcji (współczynnik korelacji 0,2). Występuje natomiast bardzo silna korelacja z temperaturą – współczynnik korelacji 0,9. W wyniku analizy korelacji niektóre zmienne zostały usunięte z modeli w dalszych etapach, ponadto algorytmy uczenia maszynowego samodzielnie dobierały zmienne w trakcie budowy modeli. Podstawowe statystki opisowe wszystkich zmiennych (z wyjątkiem roku i miesiąca) zawarte są w tabeli 1.

Tabela 1. Podstawowe statystyki opisowe zmiennych ilościowych
Table 1. Basic descriptive statistics of quantitative variables

	Jednostka	Średnia	Mediana	Odchylenie standardowe	Minimum	Maksimum
Polska	TJ	52 927,2	51 508,0	15 314,3	29 754,0	90 784,0
Niemcy	TJ	280 962,3	271 178,0	100 904,6	126 201,0	519 716,0
Hiszpania	TJ	108 687,5	106 817,0	19 724,8	77 373,0	159 409,0
Francja	TJ	145 546,9	130 988,0	73 931,2	44 833,0	307 937,0
Włochy	TJ	236 227,7	199 531,0	84 130,9	115 557,0	424 472,0
Holandia	TJ	126 597,4	118 252,0	44 420,1	67 073,0	244 402,0
Wielka Brytania	TJ	274 779,8	270 586,0	88 835,1	133 719,0	495 290,0
Norwegia	TJ	17 149,1	16 354,0	12 250,6	721,0	76 692,0
Turcja	TJ	142 012,1	130 968,0	37 700,6	89 247,0	249 004,0
Cena	EUR/mln BTU	3,0	2,7	1,1	1,5	8,2
Wynagrodzenie	PLN	3 784,1	3 775,8	474,0	2 830,4	4 973,7
Produkcja	%	91,9	91,5	11,7	67,0	120,0
Średnia temperatura	°C	4,7	4,0	6,9	−11,9	16,6

Do oceny trafności prognoz otrzymanych poszczególnymi metodami użyto miar błędów prognoz ex post: pierwiastka błędu średniokwadratowego prognozy RMSE (ang. *root mean square error*):

$$\mathrm{RMSE} = \left[\frac{1}{T-t}\sum_{i=t+1}^{T}(y_i - y_i^*)^2\right]^{0,5}$$

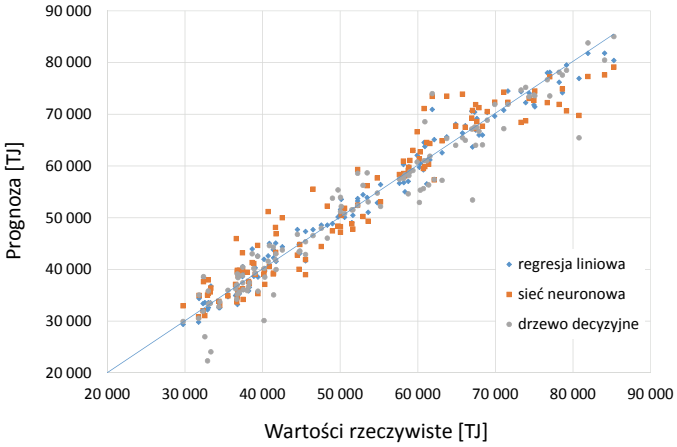
oraz średniego bezwzględnego błędu procentowego prognozy MAPE (ang. *mean absolute percentage error*)
gdzie: $[t+1, T]$ jest przedziałem czasowym, y_i jest wartością rzeczywistą w czasie t oraz y_i^* oznacza prognozę w czasie t.

$$\mathrm{MAPE} = \frac{1}{T-t}\sum_{i=t+1}^{T}\frac{|y_i - y_i^*|}{y_i}$$

W pierwszym etapie wykorzystano wszystkie dane bez ich wstępnego przetworzenia (*preprocessingu*) oraz przy użyciu automatycznego losowego podziału na zbiór uczący (70% danych) i testowy (30% danych). Modele zbudowane zostały w oparciu o zbiór uczący. Przy średniej wartości miesięcznego zużycia gazu wynoszącej 52 927 TJ – błędy RMSE i MAPE dla poszczególnych modeli wyliczone dla wszystkich posiadanych danych wynoszą odpowiednio:

- sztuczna sieć neuronowa – RMSE 4077, MAPE 6,0%;
- drzewo decyzyjne – RMSE 3755, MAPE 5,3%;
- regresja liniowa – RMSE 2350, MAPE 3,7%.

Najniższy błąd prognozy uzyskała klasyczna regresja liniowa, a nie – jak oczekiwano – metody uczenia maszynowego. Niemniej i tak błędy prognoz wszystkich metod można uznać za niezbyt duże, co widoczne jest na rysunku 1, obrazującym relację prognoz otrzymanych poszczególnymi metodami do wartości rzeczywistego zużycia gazu. Środkowa linia wyznacza idealne prognozy.



Rys. 1. Wartości prognoz w relacji do wartości rzeczywistych
Fig. 1. Values of forecasts in relation to actual values

W drugim etapie uzupełniono brakujące dane dotyczące średniego wynagrodzenia w gospodarce za lata 2008–2009 prognozami pochodzącymi z modelu Wintersa. Za pomocą średnich ruchomych uzupełniono także pojedyncze braki w danych dotyczące zużycia gazu w niektórych krajach. Sprawdzono również występowanie odstających danych (ang. *outliers*), za które uznane byłyby dane o wartościach wyższych lub niższych od średniej o trzy odchylenia standardowe. Nie znaleziono takich danych dla żadnej ze zmiennych. Tak przygotowanych danych użyto w procesie automatycznej budowy modeli prognostycznych analogicznie do etapu pierwszego. Otrzymano ponownie trzy modele, dla których błędy prognoz *ex post* dla wszystkich danych wynoszą odpowiednio:

- sztuczna sieć neuronowa – RMSE 3821, MAPE 5,6%;
- drzewo decyzyjne – RMSE 1249, MAPE 1,7%;
- regresja liniowa – RMSE 2310, MAPE 3,6%.

Tym razem najwyższą trafnością prognoz wykazał się algorytm drzewa decyzyjnego. Poza tym wszystkie modele poprawiły nieznacznie swoją trafność.

Podstawowym problemem w przypadku użycia modeli uczenia maszynowego w prognozowaniu jest to, że budowa modeli standardowo opiera się na losowym zbiorze uczącym, podczas gdy prognoza powinna zostać utworzona na podstawie wcześniejszego okresu dla okresu następnego – prognozowanego. W związku z tym w trzecim etapie podzielono dane na zbiór uczący utworzony z danych miesięcznych za okres styczeń 2008 – luty 2016 oraz zbiór testowy zawierający dane z okresu marzec 2016 – luty 2018. Na podstawie pierwszego zbioru utworzono modele prognostyczne, które następnie zweryfikowano na zbiorze testowym. Błędy prognoz *ex post* kolejnych trzech modeli dla zbioru testowego (okresu prognozowanego) wynoszą odpowiednio:

- sztuczna sieć neuronowa – RMSE 4612, MAPE 5,5%;
- drzewo decyzyjne – RMSE 5988, MAPE 8,3%;
- regresja liniowa – RMSE 3603, MAPE 4,8%.

W trzecim etapie najwyższą trafność osiągnęła regresja liniowa, od której nieznacznie gorsza okazała się sieć neuronowa. Drzewo decyzyjne uzyskało wyraźnie gorszą trafność od pozostałych dwóch algorytmów. Wynika z tego, że algorytm drzewa decyzyjnego zaimplementowany w MS SQL Server jest podatny na zjawisko przetrenowania w procesie uczenia (nadmiernego dopasowania modelu do próby). Najlepsza trafność prognoz drzewa decyzyjnego uzyskana dla zbioru wszystkich danych nie została powtórzona dla wyodrębnionego zbioru testowego w okresie prognozy.

W literaturze przedmiotu przyjmuje się, że normalizacja zmiennych poprawia trafność predykcji modeli, w szczególności sztucznych sieci neuronowych. W związku z tym w czwartym etapie dokonano początkowo normalizacji zmiennych metodą standaryzacji *Z-score* zgodnie ze wzorem:

$$x_{z-score} = \frac{x - E(x)}{\sqrt{Var(x)}}$$

gdzie *x* oznacza wartość zmiennej.

Okazało się jednak, że standaryzacja nie zmieniła wyników prognoz otrzymanych wszystkimi analizowanymi metodami, dlatego dalsze obliczenia i wyniki związane ze standaryzacją zmiennych zostały pominięte w artykule.

W kolejnym kroku dokonano normalizacji danych metodą min-max, przydatnej szczególnie w odniesieniu do uczenia sieci neuronowych, ponieważ ogranicza zakres zmiennych wejściowych do przedziału [0, 1], wymaganego w przypadku zastosowania sigmoidalnej funkcji aktywacji neuronów. Standaryzacja min-max wyrażona jest następującym wzorem:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Błędy prognoz ex post kolejnych trzech modeli utworzonych dla danych znormalizowanych dla zbioru testowego (okresu prognozowanego) wynoszą odpowiednio:

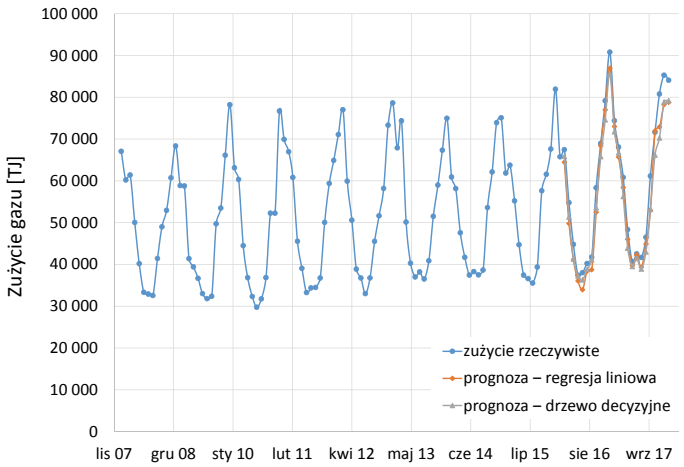
- sztuczna sieć neuronowa – RMSE 2863, MAPE 9,5%;
- drzewo decyzyjne – RMSE 2595, MAPE 6,8%;
- regresja liniowa – RMSE 1625, MAPE 4,8%.

Zużycie gazu przez kraje zachodniej i centralnej Europy ze względu na zbliżone warunki klimatyczne wykazuje się wysoką korelacją – współczynniki korelacji liniowej Pearsona dla Niemiec, Holandii, Francji, Włoch i Wielkiej Brytanii wynoszą od 0,90 do 0,95. Prowadzi to do współliniowości zmiennych dotyczących zużycia gazu przez wymienione kraje, co potwierdzone zostało wysokimi wartościami współczynników VIF (ang. *variance inflation factor*) – powyżej 10,0. W związku z tym w ostatnim, piątym etapie usunięto z modeli dane dotyczące wymienionych krajów z wyjątkiem Niemiec. W efekcie otrzymano następne trzy modele dla danych znormalizowanych, dla których błędy prognoz ex post dla zbioru testowego (okresu prognozowanego) wynoszą odpowiednio:

- sztuczna sieć neuronowa – RMSE 2918, MAPE 7,3%;
- drzewo decyzyjne – RMSE 1954, MAPE 5,8%;
- regresja liniowa – RMSE 1762, MAPE 5,4%.

Uzyskane wyniki nie różnią się istotnie od wyników modeli ze wszystkimi zmiennymi. Co zaskakujące, po usunięciu współliniowych zmiennych algorytm drzewa decyzyjnego zmniejszył błędy prognozy, chociaż powinien być najmniej wrażliwy na zjawisko współliniowości zmiennych, podczas gdy regresja liniowa, będąca wrażliwą na współliniowość zmiennych, pogorszyła nieznacznie trafność prognoz. Zestawienie prognoz regresji liniowej i drzewa decyzyjnego modeli zbudowanych dla danych znormalizowanych z usuniętymi zmiennymi z rzeczywistym zużyciem gazu przedstawione jest na rysunku 2.

Przyczyną przeważnie lepszej trafności prognoz regresji liniowej jest przede wszystkim fakt zastosowania typowych zmiennych objaśniających stosowanych tradycyjnie w prognozowaniu zapotrzebowania na gaz. Nie wykorzystano tego, co może być silną stroną metod uczenia maszynowego – algorytmów pozyskujących dane z Internetu, takich jak *search volume* i analiza sentymentalna, które mogą wykrywać nadchodzące tendencje cenowe, polityczne i ekologiczne w treści dokumentów umieszczanych Internecie, a także grupowanie (klastrowanie) klientów. Ponadto trafność modeli mierzono w stosunku do niewielkiej ilości posiadanych danych, co nie pozwoliło na wykazanie skuteczności modeli uczenia maszynowego w dostosowaniu się do szoków cenowych/podażowych na rynku, zmieniających nieprzewidywalnie strukturę zużycia nośników energii w dłuższym horyzoncie czasu. Po prostu dane pochodziły ze zbyt krótkiego okresu dla tego typu analiz – 10 lat. Niemniej jednak wszystkie algorytmy wykazały się zbliżoną trafnością prognoz.



Rys. 2. Zestawienie prognoz z rzeczywistym zużyciem gazu w Polsce w latach 2008–2018

Fig. 2. Comparison of forecasts with the actual gas consumption in Poland in the years 2008–2018

W przypadku szczegółowych prognoz obszarowych pojawi się problem nowych odbiorców i terenów wcześniej niezgazyfikowanych, dla których nie będzie danych historycznych. W takiej sytuacji wzorce zapotrzebowania na gaz pochodzące z wcześniejszych okresów muszą zostać uzupełnione o dane dotyczące przyszłej struktury i planowanego zużycia gazu przez odbiorców.

Podsumowanie

Wybrane modele uczenia maszynowego wykazały się dobrą trafnością prognoz miesięcznego zużycia gazu, jednakże nieznacznie ustąpiły klasycznej regresji liniowej, ze względu na

zbyt wąski zbiór zmiennych objaśniających. Błędy prognoz *ex post* dla wszystkich modeli były niewielkie i w ujęciu względnym wynosiły kilka procent, co można uznać za dobry wynik. Dla zwiększenia trafności prognoz wskazane było odpowiednie dobranie próby uczącej oraz wstępne przygotowanie danych (*preprocessing*). Niemniej jednak automatyczne narzędzia budowy modeli zapewniają dobrą trafność prognoz bez wstępnego przygotowania danych, co pozwala na oszczędność czasu i nie wymaga udziału wyspecjalizowanych analityków danych. Oprogramowanie tego typu może działać na zasadzie „czarnej skrzynki", czyniąc metody uczenia maszynowego wygodnym narzędziem budowy prognoz.

Metody uczenia maszynowego będą mogły wykazać się wyższą skutecznością wraz ze wzrostem liczby danych oraz rozszerzeniem zbioru potencjalnych zmiennych objaśniających, w tym w szczególności po zastosowaniu jakościowej analizy tekstów zawartych w Internecie. Metody uczenia maszynowego nadają się przede wszystkich do dużych zbiorów danych, nieoczyszczonych i nieprzygotowanych wstępnie do klasycznej analizy regresji, wymagającej usunięcia odstających danych, brakujących danych i skorelowanych zmiennych oraz wieloetapowej selekcji zmiennych objaśniających. W morzu danych metody uczenia maszynowego w połączeniu z narzędziami automatycznego pozyskania danych (ETL) są w stanie skutecznie tworzyć modele prognostyczne bez konieczności żmudnej ingerencji analityka we wstępne przygotowanie danych oraz dobór zmiennych objaśniających (Han et al., 2012; Šebalj et al., 2017). Pozwalają także na dowolnie częstą aktualizację postaci modeli prognostycznych, nawet po każdym uzupełnieniu zbioru danych.

Artykuł został opracowany na podstawie referatu wygłoszonego na Międzynarodowej Konferencji Naukowo-Technicznej GEOPETROL 2018 pt.: *Rozwój technik poszukiwania i eksploatacji złóż węglowodorów*. Zakopane-Kościelisko, 17–20.09.2018 r.

Literatura

Balfe P., Kelp O., 2014. Gas consumption Forecasting. A methodology. Sydney: Acil Allen Consulting.

Bank Danych Lokalnych, b.d. <https://bdl.stat.gov.pl/BDL/dane/podgrup/temat> (dostęp: czerwiec 2018).

Bank Danych Makroekonomicznych, b.d. <https://bdm.stat.gov.pl/> (dostęp: czerwiec 2018).

Bianco V., Scarpa F., Tagliafico L., 2014a. Analysis and future outlook of natural gas consumption in the Italian residential sector. *Energy Conversion and Management*, 87: 754–764. DOI: 10.1016/j.enconman.2014.07.081.

Bianco V., Scarpa F., Tagliafico L., 2014b. Scenario analysis of non-residential natural gas consumption in Italy. *Applied Energy*, 113: 392–403. DOI: 10.1016/j.apenergy.2013.07.054.

Eurostat, b.d. Database. <http://ec.europa.eu/eurostat/web/energy/data/database> (dostęp: czerwiec 2018).

Freemeteo, b.d. <https://freemeteo.pl/> (dostęp: czerwiec 2018).

Gaweł B., Rębiasz B., Skalna I., 2016. Data mining methods for long-term forecasting of market demand for industrial goods. W: Z. Wilimowska et al. (eds.), Information Systems Architecture and Technology: Proceedings of the 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV (s. 3–13). Switzerland: Springer International Publishing. DOI: 10.1007/978-3-319-28567-2.

Han J., Kamber M., Pei J., 2012. Data Mining: Concepts and Techniques. San Francisco: Elsevier.

Holdaway K., 2014. Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data Driven Models. New Jersey: John Wiley & Sons.

IndexMundi, b.d. <https://www.indexmundi.com> (dostęp: czerwiec 2018).

Larose D., 2006. Odkrywanie wiedzy z danych. Warszawa: PWN.

Microsoft, b.d.-a. Omówienie programu Visual Studio 2017. <https://docs.microsoft.com/en-us/visualstudio/ide/visual-studio-ide> (dostęp: czerwiec 2018).

Microsoft, b.d.-b. Edycje programu SQL Server 2017. <https://www.microsoft.com/pl-pl/sql-server/sql-server-2017-editions> (dostęp: czerwiec 2018).

Paliński A., 2018. Hurtownie danych i eksploracja danych w prognozowaniu popytu na gaz i usługi magazynowania gazu. *Nafta-Gaz*, 4: 283–289. DOI: 10.18668/NG.2018.04.04.

Sánchez-Úbeda E., Berzosa A., 2007. Modeling and forecasting industrial end-use natural gas consumption. *Energy Economics*, 29(4): 710–742. DOI: 10.1016/j.eneco.2007.01.015.

Šebalj D., Mesarić J., Dujak D., 2017. Predicting Natural Gas Consumption – A Literature Review. W: Proceedings of Central European Conference on Information and Intelligent Systems (s. 293–300). Varaždin, Croatia.

Suykens J. et al., 1996. Modelling the Belgian gas consumption using neural networks. *Neural Processing Letters*, 4(3): 157–166. DOI: 10.1007/BF00426024.

Dr hab. inż. Andrzej PALIŃSKI
Adiunkt na Wydziale Zarządzania
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie,
ul. Gramatyka 10
30-067 Kraków
E-mail: *palinski@zarz.agh.edu.pl*