Wiadomości Statystyczne. The Polish Statistician, 2020, vol. 65, 10, 24–48 DOI: 10.5604/01.3001.0014.4315
Statystyka w praktyce / Statistics in practice
Otrzymano/received: 10.06.2019, zaakceptowano/accepted: 30.07.2020

Prognozowanie brakujących danych w szeregach czasowych przy zastosowaniu modeli hybrydowych – podejście teoretyczne i empiryczne

Jan Zawadzki[a]

Streszczenie. Celem artykułu jest przedstawienie możliwości prognozowania brakujących danych w szeregach czasowych dla danych godzinnych przy zastosowaniu modeli hybrydowych. W badaniu wykorzystano hybrydowe modele szeregu czasowego oraz modele regresyjne ze złożonymi wahaniami sezonowymi. Wahania złożone dla danych godzinnych mogą być sumą lub iloczynem wahań o cyklach rocznym, tygodniowym i dobowym, a wahania o długości cyklu wyrażonej liczbą parzystą (12-miesięczne i 24-godzinne) mogą być opisywane za pomocą regularnych modeli hierarchicznych.

Rozważania o charakterze teoretycznym podparto analizą empiryczną zapotrzebowania na energię elektryczną w okresach godzinnych w wybranej aglomeracji. Wykorzystano dane statystyczne obejmujące trzy kolejne lata pierwszej dekady XXI w., które zostały udostępnione przez przedsiębiorstwo zajmujące się dystrybucją energii elektrycznej i włączone do Banku Danych Katedry Zastosowań Matematyki w Ekonomii Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie. Przyjęto założenie, że luki niesystematyczne będą występowały w przypadku każdego rodzaju wahań sezonowych. Otrzymane wyniki wskazują na użyteczność modeli hybrydowych w prognozowaniu zjawisk ekonomicznych o bardzo dużej częstotliwości prowadzenia obserwacji.

Słowa kluczowe: dane godzinne, sezonowość złożona, modele hybrydowe, luki niesystematyczne

JEL: C32, C53

Forecasting missing data in time series with the application of hybrid models – theoretical and empirical approach

Abstract. The main goal of the article is to present the possibilities of forecasting missing observations in time series for hourly data with the application of hybrid models. Hybrid time series models and regression models with complex seasonal fluctuations were used in the study. Complex fluctuations for hourly data can be either a sum or a product of fluctuations of annual, weekly and daily cycles, while fluctuations the length of a cycle expressed by an even number (12-month and 24-hour ones) can be described using regular hierarchical models.

The theoretical considerations were illustrated by an empirical analysis of the demand for electricity in hourly periods in a selected agglomeration. The statistical data covered three consecutive years of the first decade of the 2000s. The data were provided by an electricity distribution company and included in the Data Bank of the Department of Applied Mathematics in Economics of the West Pomeranian University of Technology in Szczecin. It was assumed

[a] Zachodniopomorski Uniwersytet Technologiczny w Szczecinie, Wydział Ekonomiczny.
ORCID: https://orcid.org/0000-0002-2996-3180.

© Główny Urząd Statystyczny

that non-systematic gaps occur with regard to all types of seasonal fluctuations. The obtained results indicate the usefulness of hybrid models in forecasting economic phenomena subject to very frequent observations.

Keywords: hourly data, complex seasonality, hybrid models, non-systematic gaps

1. Wprowadzenie teoretyczne

W modelowaniu i prognozowaniu zmiennych dla danych godzinnych znajdują zastosowanie klasyczne modele szeregu czasowego z potrójnie złożonymi wahaniami sezonowymi w postaci addytywnej lub multiplikatywnej, w których każdy rodzaj wahań jest opisywany za pomocą osobnych podzbiorów zmiennych zero-jedynkowych.

Zapis analityczny modelu szeregu czasowego wykazującego złożoną sezonowość w postaci addytywnej z trendem liniowym jest następujący (Kufel, 2010; Szmuksta-Zawadzka i Zawadzki, 2011):

$$Y_t = \alpha_1 t + \alpha_0 + \sum_{i=1}^{12} b_{0i} M_{it} + \sum_{j=1}^{7} c_{0j} D_{jt} + \sum_{k=1}^{24} d_{0k} G_{kt} + U_t \tag{1}$$

$$t = 1, 2, \dots, n$$

przy warunkach

$$\sum_{i=1}^{12} b_{0i} = \sum_{j=1}^{7} c_{0j} = \sum_{k=1}^{24} d_{0k} = 0,$$

gdzie:

M_{it} – zmienne zero-jedynkowe przyjmujące wartość 1 w i-tym miesiącu (i = 1, 2, ..., 12),

D_{jt} – zmienne zero-jedynkowe przyjmujące wartość 1 w j-tym dniu tygodnia (j = 1, 2, ..., 7),

G_{kt} – zmienne zero-jedynkowe przyjmujące wartość 1 w k-tej godzinie dnia tygodnia (k = 1, 2, ..., 24),

U_t – składnik losowy.

Model w postaci wykładniczej wyraża się wzorem

$$Y_t = \exp\left(\alpha_1 t + \alpha_0 + \sum_{i=1}^{12} b_{0i} M_{it} + \sum_{j=1}^{7} c_{0j} D_{jt} + \sum_{k=1}^{24} d_{0k} G_{kt} + U_t\right) \tag{2}$$

$$t = 1, 2, \dots, n$$

przy warunkach

$$\sum_{i=1}^{12} b_{0i} = \sum_{j=1}^{7} c_{0j} = \sum_{k=1}^{24} d_{0k} = 0.$$

Objaśnienia symboli są takie same jak w modelu (1).

Po obustronnym zlogarytmowaniu za pomocą logarytmu naturalnego otrzymujemy jego postać zlinearyzowaną względem $\ln Y_t$.

W literaturze coraz częściej można spotkać przykłady wykorzystania modeli hierarchicznych z wahaniami sezonowymi w modelowaniu i prognozowaniu na podstawie kompletnych szeregów czasowych. Zazwyczaj dotyczy to szeregów dla danych miesięcznych i ewentualnie dziennych (m.in. Athanasopoulos i in., 2017; Hyndman i in., 2011), rzadziej dla danych godzinnych czy półgodzinnych (np. Kwac i in., 2014; Taieb i in., 2017).

Modele, w których przynajmniej jeden rodzaj wahań o długości cyklu wyrażonej liczbą parzystą jest opisywany za pomocą regularnych modeli hierarchicznych, noszą nazwę modeli hybrydowych (mieszanych). W modelach tych wahania o długości cyklu wyrażonej liczbą nieparzystą (7-dniowe) są opisywane tak jak w modelu (1), za pomocą zmiennych zero-jedynkowych. Metody i przykłady ich zastosowań można znaleźć m.in. w pracach Szmuksta-Zawadzkiej i Zawadzkiego (2014a) oraz Zawadzkiego (2018). Regularne modele hierarchiczne mogą być wykorzystywane do opisu wahań o parzystej długości cykli dla danych godzinnych (w analizowanym przypadku – o cyklach rocznym i dobowym).

Modele, w których przynajmniej jeden rodzaj wahań jest opisywany za pomocą modelu hierarchicznego, to pojedyncze modele hybrydowe. Jeżeli wykorzystuje się je do opisu obu rodzajów wahań, nazywa się je podwójnymi modelami hybrydowymi. Wahania o cyklu tygodniowym (nieparzystym) każdorazowo są opisywane za pomocą zmiennych zero-jedynkowych D_{jt}.

Definicja regularnego modelu hierarchicznego zamieszczona została w pracy Szmuksta-Zawadzkiej i Zawadzkiego (2002). Wynika z niej, że jest to model o podzielnikach p_i cyklu wahań okresowych (sezonowych) m, będącego liczbą parzystą, spełniający jednocześnie dwa warunki:

$$2 \leq p_i \leq \frac{m}{2} \quad \text{oraz} \quad \prod_i p_i = m. \tag{3}$$

Modele hierarchiczne dla wahań o cyklu 12-miesięcznym i 24-godzinnym są oznaczane odpowiednio jako HM i HG. Liczby występujące po tych symbolach oznaczają kolejne podzielniki długości cyklu wahań.

Liczba regularnych modeli hierarchicznych dla danego cyklu wahań m jest równa liczbie permutacji oraz permutacji z powtórzeniami podzielników p_i. Dla danych miesięcznych o cyklu rocznym ($m = 12$) jest ich siedem:

- cztery dwustopniowe (HM26, HM34, HM43, HM62);
- trzy trzystopniowe (HM223, HM232, HM322).

Dla danych o cyklu dobowym ($m = 24$) jest ich 19:

- sześć dwustopniowych (HG2_12, HG38, HG46, H64, HG83, HG12_2);
- dziewięć trzystopniowych (HG226, HG234, HG243, HG262, HG342, HG423, HG432, HG226, HG622);
- cztery czterostopniowe (HG2223, HG2232, HG2322, HG3222).

Liczba szacowanych podwójnych modeli hybrydowych jest równa iloczynowi liczb modeli pojedynczych i wynosi 7 · 19 = 133. Dla jednej postaci analitycznej szacowanych jest zatem łącznie 7 + 19 + 133 = 159 modeli hybrydowych. Do tego należy doliczyć jeden model klasyczny (z trzema zespołami zmiennych zero-jedynkowych) dany równaniem (1), służący jako punkt odniesienia dla modeli hybrydowych. Modele te pozwalają na „wyrównywanie" obserwacji w okresach będących skutkiem zakłóceń losowych lub zakłóceń występujących w okresach, w których istnieją luki w danych.

Poniżej podano zapisy analityczne dwu-, trzy- i czterostopniowych liniowych modeli z trendem liniowym (Szmuksta-Zawadzka i Zawadzki, 2014a).

Zapis dwustopniowego modelu hierarchicznego wygląda następująco:

$$Y_t = \alpha_1 t + \alpha_0 + \sum_{s=1}^{p_1} b_{0s} Q_{st} + \sum_{r=1}^{p_2} b_{0sr} Q_{srt} + U_{srt} \tag{4}$$

$$t = 1, 2, \dots, n$$

przy warunkach

$$\sum_{s=1}^{p_1} b_{0s} = \sum_{r=1}^{p_2} b_{0sr} = 0.$$

Model trzystopniowy wyraża się wzorem

$$Y_t = \alpha_1 t + \alpha_0 + \sum_{s=1}^{p_1} b_{0s} Q_{st} + \sum_{r=1}^{p_2} b_{0sr} Q_{srt} + \sum_{l=1}^{p_3} b_{0srlt} Q_{srlt} + U_{srlt} \tag{5}$$

$$t = 1, 2, \dots, n$$

przy warunkach

$$\sum_{s=1}^{p_1} b_{0s} = \sum_{r=1}^{p_2} b_{0sr} = \sum_{l=1}^{p_3} b_{0srlt} = 0.$$

Model czterostopniowy można zapisać jako

$$Y_t = \alpha_1 t + \alpha_0 + \sum_{s=1}^{p_1} b_{0s} Q_{st} + \sum_{r=1}^{p_2} b_{0sr} Q_{srt} + \sum_{l=1}^{p_3} b_{0srlt} Q_{srlt} + \\ + \sum_{k=1}^{p_4} b_{0srlkt} Q_{srlkt} + U_{srlkt} \qquad t = 1, 2, \dots, n \tag{6}$$

przy warunkach

$$\sum_{s=1}^{p_1} b_{0s} = \sum_{r=1}^{p_2} b_{0sr} = \sum_{l=1}^{p_3} b_{0srlt} = \sum_{k=1}^{p_4} b_{0srlkt} = 0.$$

Postać zmiennych Q zależy od liczby stopni hierarchii, liczby i kolejności podzielników modelu oraz długości cyklu wahań. Występujące we wzorach (4)–(6) zmienne zero-jedynkowe Q dla cyklu 24-godzinnego odnoszą się do 19 modeli dwu-, trzy- i czterostopniowych. Liczba zmiennych na każdym stopniu hierarchii, z uwagi na warunek sumowalności parametrów do 0, jest o 1 mniejsza od wartości podzielników $p_1, p_2,\ p_3$ lub p_4. Przykładowo w modelu dwustopniowym HG12_2 ich liczba wynosi odpowiednio 11 i 1. Zmienne zero-jedynkowe Q_{st} ($s = 1, 2, ..., 11$) przyjmują wartości równe 1 dla każdego okresu dwugodzinnego: $Q_{1t} = 1$ dla godziny 1 i 2 oraz 0 dla pozostałych godzin cyklu dobowego, Q_{2t} dla 3 i 4, ..., Q_{11t} dla 21 i 22. Natomiast zmienna Q_{srt} ($r = 1, 2$) jest tylko jedna, ponieważ $p_2 - 1 = 1$. Zmienna Q_{s1t} przyjmuje wartość 1 dla pierwszej (nieparzystej) godziny każdego okresu dwugodzinnego.

W siedmiu dwu- i trzystopniowych modelach hierarchicznych opisujących wahania o cyklu rocznym zmienne Q przyjmują inne wartości. Przykładowo dla modelu HM43 podzielniki wynoszą: $p_1 = 4$ i $p_2 = 3$. Liczby zmiennych Q są o 1 mniejsze. Zmienne zero-jedynkowe $Q_{st(M)}$ ($s = 1, 2, 3$) przyjmują wartości 1 w s-tym kwartale i 0 w pozostałych, a zmienne $Q_{srt(M)}$ ($r = 1, 2$) – wartości 1 w r-tych miesiącach każdego kwartału. Wszystkie omawiane w pracy modele wykazujące złożoną sezonowość: klasyczne szeregu czasowego, regresyjne oraz hierarchiczne (hybrydowe) są modelami jednorównaniowymi.

Idea budowy i konwencja zapisu regularnych modeli hierarchicznych zaproponowana przez Szmuksta-Zawadzką i Zawadzkiego (2002) dla danych miesięcznych (roz-

szerzona potem na dane dekadowe, dzienne i godzinne) jest koncepcją uniwersalną i odnosi się do modeli opisujących wahania sezonowe o parzystej długości cykli, zarówno niezłożonych, jak i złożonych. Jest ona podobna do prezentowanej m.in. w pracy Athanasopoulosa i in. (2017) dla danych miesięcznych. Podobieństwo dotyczy agregacji na okresy dwu-, trzy-, cztero- i sześciomiesięczne. Różnica, oprócz sposobu zapisu, polega natomiast na tym, że na najniższym stopniu hierarchii liczy się średnie dla okresów jednoimiennych. Przykładowo w przypadku wahań o okresie 24-godzinnym w modelu HG12_2 liczy się średnie dla godzin parzystych i nieparzystych.

W modelach hierarchicznych parametry strukturalne są uśrednionymi parametrami klasycznych modeli szeregu czasowego ze zmiennymi zero-jedynkowymi. Sposób uśredniania zależy od liczby stopni hierarchii oraz liczby i kolejności podzielników.

Dla cyklu 24-godzinnego w modelach (1) i (2), przy uwzględnieniu warunku sumowalności do 0, szacowane są 23 parametry d_{0k} ($k = 1, 2, ..., 23$). Ocenę parametru d_{024} oblicza się jako sumę 23 parametrów i nadaje się jej przeciwny znak. Natomiast w dwustopniowym modelu hierarchicznym HG12_2 szacowanych jest 11 parametrów b_{0s} ($s = 1, 2, ..., 11$) na pierwszym stopniu hierarchii i 1 parametr na drugim stopniu – b_{0s1}. Parametry b_{0s} są średnimi z parametrów d_{0k} dla kolejnych okresów dwugodzinnych: $b_{01} = 1/2 \cdot (d_{01} + d_{02})$; $b_{02} = 1/2 \cdot (d_{03} + d_{04})$; ...; $b_{011} = 1/2 \cdot (d_{021} + d_{022})$. Parametr b_{012} wyznacza się z warunku sumowalności do 0. Jedynie parametr b_{0s1} jest szacowany bezpośrednio jako średnia z parametrów d_{0k} dla godzin nieparzystych: $b_{0s1} = 1/12 \cdot (d_{01} + d_{03} + d_{05} + \ ... \ + d_{023})$. Z warunku sumowalności do 0 wynika, że $b_{0s2} = -b_{0s1}$.

Literatura poświęcona metodom prognozowania brakujących danych w szeregach czasowych jest dość obszerna, jednak dominują prace dotyczące prognozowania dla okresów jednostkowych o cyklu co najwyżej tygodniowym. Spośród nich należy przywołać monografie o charakterze metodologicznym autorstwa Allisona (2001), Little'a i Rubina (2002) oraz Raudenbusha i Bryka (2002). Prace podejmujące problematykę prognozowania brakujących danych w godzinnych szeregach czasowych lub dla okresów jednostkowych krótszych niż godzina należą raczej do rzadkości i na ogół odnoszą się do prognozowania zjawisk meteorologicznych. Przykładowo można tu wymienić pracę Henna i in. (2013), poświęconą prognozowaniu temperatury. Weron (2006) rozpatruje prognozowanie w szeregach z lukami, w których zapotrzebowanie na energię jest mierzone w odstępach 5-minutowych. Metoda polega na zastępowaniu brakujących danych średnimi z obserwacji z dłuższych odcinków czasu, z sąsiadujących obserwacji lub np. z poprzedniego tygodnia, po wcześniejszym wyeliminowaniu obserwacji odstających. Badania nad zastosowaniami metod prognozowania brakujących danych w szeregach czasowych z potrójnie złożonymi wahaniami sezonowymi (dla danych godzinnych) były prowadzone do roku 2016 przez Szmuksta-Zawadzką i Zawadzkiego (2011, 2014a, 2015), a następnie przez Zawadzkiego (2018).

Celem niniejszego artykułu jest przedstawienie możliwości prognozowania brakujących danych w szeregach czasowych dla danych godzinnych przy zastosowaniu

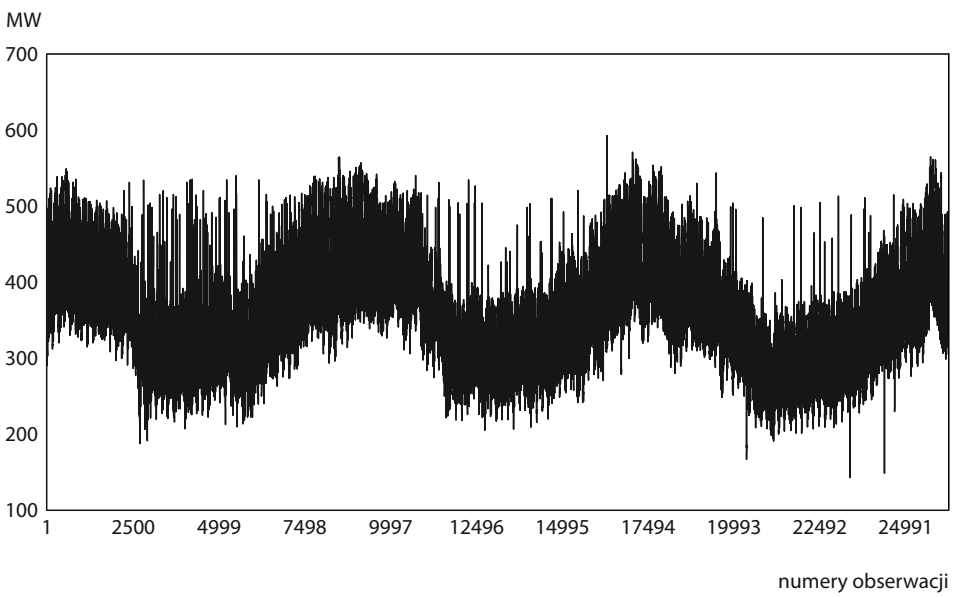
modeli hybrydowych. W omawianym badaniu wykorzystano hybrydowe modele szeregu czasowego oraz modele regresyjne[1] ze złożonymi wahaniami sezonowymi.

2. Metoda badania

Modelowaniu, a następnie prognozowaniu poddano zapotrzebowanie na energię elektryczną w okresach godzinnych w aglomeracji A. Dane statystyczne o kształtowaniu się zapotrzebowania na moc zostały udostępnione przez przedsiębiorstwo zajmujące się dystrybucją energii elektrycznej i włączone do Banku Danych Katedry Zastosowań Matematyki w Ekonomii Zachodniopomorskiego Uniwersytetu Technologicznego (ZUT) w Szczecinie. Odnoszą się one do trzech kolejnych lat pierwszej dekady XXI w.[2] Szereg wyjściowy (bez luk) obejmował dwa lata, tj. $2 \cdot 365 \cdot 24 = 17520$ obserwacji ($t = 1, 2, ..., 17520$). Trzeci rok był okresem empirycznej weryfikacji prognoz i zawierał $365 \cdot 24 = 8760$ obserwacji ($t = 17521, ..., 26280$).

Kształtowanie się zmiennej w okresie estymacyjnym i prognozowanym zostało przedstawione na wykr. 1.

Wykr. 1. Zapotrzebowanie na moc elektryczną



Źródło: Bank Danych Katedry Zastosowań Matematyki w Ekonomii ZUT w Szczecinie.

[1] Określenie *modele regresyjne* przyjęto m.in. za pracą Witkowskiej i in. (2000).

[2] Warunkiem udostępnienia danych było zobowiązanie do zachowania anonimowości firmy i nieujawniania konkretnych lat, których dotyczą.

W pracy Zawadzkiego (2018) rozpatrywano wariant, w którym luki niesystematyczne występowały w ośmiu godzinach: 2, 6, 8, 10, 14, 16, 18, 22, w okresie od 1 lipca pierwszego roku do 30 czerwca drugiego roku (wariant G). Łączna ich liczba wynosiła 2920 (365 · 8). W badaniu omawianym w artykule założono, że luki niesystematyczne, oprócz wymienionych godzin, występują także w dwóch dniach każdego z 52 tygodni (wtorki i czwartki) i każdego z 62 dni dwóch miesięcy (sierpień i październik) – wariant GDM. W wyznaczeniu liczby luk dla tygodni należy wziąć pod uwagę fakt ich wcześniejszego uwzględnienia dla godzin. W każdym z dwóch dni tygodnia liczba luk zwiększy się o 16 godzin. Otrzymamy zatem 52 · 2 · 16 godzin = = 1664 godziny. Natomiast w przypadku wymienionych miesięcy należy uwzględnić występowanie luk dla godzin i tygodni. Przyrost liczby luk wyniesie więc (62 – 19) · 16 = 688 godzin. Oznacza to zwiększenie łącznej liczby luk w tym wariancie z 2920 do 5272, tj. o ponad 80%. Luki otrzymano przez wymazanie odpowiednich danych z pełnego szeregu.

Rozmieszczenie luk osobno w przypadku ośmiu godzin dla każdego z 365 dni, w przypadku dwóch dni tygodnia dla każdego z 52 tygodni i w przypadku 62 dni dla dwóch miesięcy przedstawiono w zestawieniach 1–3. Puste komórki oznaczają występowanie luk, a przez x oznaczono dostępne dane. Występowanie pustych komórek przynajmniej w jednej kolumnie oznacza luki systematyczne.

Zestawienie 1. Układ luk dla danych godzinnych

Lata i półrocza	Godziny																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
I: 1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2	x		x	x	x		x		x		x	x	x		x		x		x	x	x		x	x
II: 1	x		x	x	x		x		x		x	x	x		x		x		x	x	x		x	x
2	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Źródło: opracowanie własne.

Zestawienie 2. Układ luk dla dni tygodnia

Lata i półrocza	Dni tygodnia						
	poniedziałek	wtorek	środa	czwartek	piątek	sobota	niedziela
I: 1	x	x	x	x	x	x	x
2	x		x		x	x	x
II: 1	x		x		x	x	x
2	x	x	x	x	x	x	x

Źródło: opracowanie własne.

Zestawienie 3. Układ luk dla miesięcy

Lata i półrocza	Miesiące											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
I: 1	x	x	x	x	x	x	x	x	x	x	x	x
2	x	x	x	x	x	x	x		x		x	x
II: 1	x	x	x	x	x	x	x		x		x	x
2	x	x	x	x	x	x	x	x	x	x	x	x

Źródło: opracowanie własne.

Do modelowania i prognozowania wykorzystano pojedyncze i podwójne hybrydowe modele szeregu czasowego i regresyjne. Hybrydowe modele regresyjne, będące rozbudowanymi modelami szeregu czasowego, zawierają opóźnioną o 24 godziny zmienną endogeniczną oraz zmienne zero-jedynkowe opisujące występowanie świąt i jednego dnia przedświątecznego (Wielkiej Soboty). Wprowadzenie dni świątecznych uzasadnia się znacznie niższym poziomem zapotrzebowania na energię elektryczną w tych dniach (zob. Witkowska i in., 2000).

Poniżej podano zapisy analityczne dwu-, trzy- i czterostopniowych hierarchicznych liniowych modeli regresyjnych. Zapis dwustopniowego regresyjnego modelu hierarchicznego jest następujący:

$$Y_t = \alpha_1 t + \alpha_0 + \sum_{s=1}^{p_1} b_{0s} Q_{st} + \sum_{r=1}^{p_2} b_{0sr} Q_{srt} + \beta Y_{t-24} + \sum_{z=1}^{10} \gamma_z S_{zt} + U_{srlt} \tag{7}$$

$$t = 1, 2, \dots, n$$

przy warunkach

$$\sum_{s=1}^{p_1} b_{0s} = \sum_{r=1}^{p_2} b_{0sr} = 0.$$

Model trzystopniowy wyraża się wzorem

$$Y_t = \alpha_1 t + \alpha_0 + \sum_{s=1}^{p_1} b_{0s} Q_{st} + \sum_{r=1}^{p_2} b_{0sr} Q_{srt} + \sum_{l=1}^{p_3} b_{0srlt} Q_{srlt} + \beta Y_{t-24} + \\ + \sum_{z=1}^{10} \gamma_z S_{zt} + U_{srlt} \qquad t = 1, 2, \dots, n \tag{8}$$

przy warunkach

$$\sum_{s=1}^{p_1} b_{0s} = \sum_{r=1}^{p_2} b_{0sr} = \sum_{l=1}^{p_3} b_{0srlt} = 0.$$

Model czterostopniowy można zapisać następująco:

$$Y_t = \alpha_1 t + \alpha_0 + \sum_{s=1}^{p_1} b_{0s} Q_{st} + \sum_{r=1}^{p_2} b_{0sr} Q_{srt} + \sum_{l=1}^{p_3} b_{0srlt} Q_{srlt} + \\ + \sum_{k=1}^{p_4} b_{0srlkt} Q_{srlkt} + \beta Y_{t-24} + \sum_{z=1}^{10} \gamma_z S_{zt} + U_{srlt} \qquad t = 1, 2, \dots, n \tag{9}$$

przy warunkach

$$\sum_{s=1}^{p_1} b_{0s} = \sum_{r=1}^{p_2} b_{0sr} = \sum_{l=1}^{p_3} b_{0srlt} = \sum_{k=1}^{p_4} b_{0srlkt} = 0.$$

Zmienne zero-jedynkowe S_{zt} przyjmują wartość 1 w następujące dni świąteczne: Nowy Rok, Wielkanoc, 1 i 3 Maja, Boże Ciało, Wniebowzięcie Najświętszej Marii Panny (15 sierpnia), Wszystkich Świętych (1 listopada), Dzień Niepodległości (11 listopada), Boże Narodzenie (25 grudnia) i Wielką Sobotę.

W modelach wykładniczych zamiast Y_{t-24} występuje $\ln Y_{t-24}$.

W modelach hybrydowych przynajmniej jeden rodzaj wahań o parzystej długości cyklu (24-godzinne lub/i 12-miesięczne) jest opisywany za pomocą regularnych modeli hierarchicznych.

Dla obu rodzajów modeli (szeregu czasowego i regresyjnych) szacowano także dwie postaci analityczne: liniową i wykładniczą.

Parametry strukturalne zarówno klasycznych modeli szeregu czasowego ze złożonymi wahaniami sezonowymi, dane równaniami (1) i (2), jak i modeli regresyjnych szacowano klasyczną metodą najmniejszych kwadratów, uwzględniającą warunki sumowalności parametrów do 0. Estymację wykonano na podstawie dostępnych danych, tzn. z wyłączeniem okresów, w których wystąpiły luki. Tę samą metodę zastosowano do szacowania parametrów modeli hybrydowych, zarówno pojedynczych, jak i podwójnych. Do estymacji parametrów wykorzystano pakiet Statistica 13.3, a do obliczenia prognoz oraz ich błędów względnych – arkusz kalkulacyjny Excel[3].

[3] W literaturze (zob. np. Athanasopoulos i in., 2017) można znaleźć informacje o pakietach w języku R przeznaczonych do szacowania modeli hierarchicznych i prognozowania na ich podstawie; są to pakiety thief i forecast.

Na podstawie predyktorów opartych na oszacowanych równaniach wyznaczano dwa rodzaje prognoz: interpolacyjne i ekstrapolacyjne. Dla okresów, w których wystąpiły luki, wyznaczono prognozy interpolacyjne. Otrzymano je przez podstawienie wartości zmiennych występujących w tym okresie po prawej stronie każdego z oszacowanych modeli. W omawianym badaniu w wariancie GDM (w którym luki występują w ośmiu okresach godzinnych, dwóch dniach tygodnia oraz dwóch miesiącach) liczba ta wynosi 5272. Z kolei dla okresu empirycznej weryfikacji prognoz, wybiegającego poza przedział czasowy „próby", dla $t = 17521, ..., 26280$, wyznaczono 8760 prognoz ekstrapolacyjnych *ex post*. Otrzymano je przez podstawienie wartości zmiennych występujących po prawej stronie równań w okresie prognozowanym ($t > 17520$).

Ponieważ dla obu rodzajów prognoz znane były realizacje zmiennej prognozowanej, obliczono dla nich średnie błędy względne dane wzorami (10)–(13).

3. Wyniki modelowania i prognozowania dla modeli hybrydowych szeregu czasowego

Z badań dotyczących prognozowania brakujących danych, m.in. w szeregach czasowych z podwójnie złożonymi wahaniami sezonowymi (dla danych dziennych), wynika, że najmniejsze wartości oceny błędów prognoz interpolacyjnych i ekstrapolacyjnych otrzymuje się zazwyczaj dla modeli hybrydowych o innej strukturze lub innej postaci analitycznej (zob. np. Szmuksta-Zawadzka i Zawadzki, 2014b). Inna struktura oznacza, że modele różnią się liczbą lub kolejnością podzielników względnie rodzajem modelu (pojedynczy lub podwójny). Natomiast inna postać analityczna oznacza, że np. model o najmniejszych błędach prognoz interpolacyjnych może być liniowy, a prognoz ekstrapolacyjnych – wykładniczy. Ilustruje to przykład liniowego hybrydowego predyktora pojedynczego L_GDM_HG122_0/1 i wykładniczego predyktora podwójnego W_GDM_HG122_HM43.

Do prognozowania wybrano predyktory charakteryzujące się najmniejszymi wartościami oceny względnych błędów prognoz interpolacyjnych ($MAPE_I$) bądź prognoz ekstrapolacyjnych ($MAPE_E$). Przy zastosowaniu każdego z tych kryteriów obliczono, jako wielkości pomocnicze, błędy drugiego rodzaju prognoz. Oznaczono je odpowiednio $MAPE_{I_E}$ oraz $MAPE_{E_I}$. Pozwoliło to na porównanie wartości tego samego miernika przy zastosowaniu różnych kryteriów dotyczących wyboru modelu dla celów prognozowania, które polegają na minimalizacji błędów $MAPE_I$ lub $MAPE_E$. Jeżeli kryterium stanowi minimalizacja $MAPE_I$, to drugim rodzajem prognoz są prognozy ekstrapolacyjne, dla których oblicza się miernik $MAPE_{I_E}$. Można go porównać z miernikiem $MAPE_E$ wyznaczonym przy zastosowaniu kryterium

minimalizacji błędów prognoz ekstrapolacyjnych. To samo odnosi się do prognoz interpolacyjnych MAPE_I oraz MAPE_{E_I}.

Wzory dotyczące wymienionych mierników mają postać:

$$\mathrm{MAPE}_I = \frac{1}{n_1} \sum_{t \epsilon I_{luk}} \frac{|Y_t - Y_{tp}|}{Y_t}, \tag{10}$$

$$\mathrm{MAPE}_{I_E} = \frac{1}{T} \sum_{t=n+1}^{T} \frac{|Y_t - Y_{tp}|}{Y_t}, \tag{11}$$

$$\mathrm{MAPE}_E = \frac{1}{T} \sum_{t=n+1}^{T} \frac{|Y_t - Y_{tp}|}{Y_t}, \tag{12}$$

$$\mathrm{MAPE}_{E_I} = \frac{1}{n_1} \sum_{t \epsilon I_{luk}} \frac{|Y_t - Y_{tp}|}{Y_t}, \tag{13}$$

gdzie:

Y – realizacje zmiennej prognozowanej,

Y_{tp} – prognozy,

n_1 – liczba brakujących danych (prognoz interpolacyjnych),

$t \epsilon I_{luk}$ – t należy do okresów, w których występują luki,

n – liczba obserwacji dla pełnego szeregu (bez luk),

T – liczba prognoz ekstrapolacyjnych.

Wielkości wybranych parametrów struktury stochastycznej dla najlepszych pojedynczych i podwójnych modeli hybrydowych szeregu czasowego w postaci liniowej i wykładniczej charakteryzujących się najmniejszymi wartościami oceny błędów MAPE_I zestawiono w tabl. 1. W ostatniej kolumnie podano oceny błędów prognoz ekstrapolacyjnych otrzymanych przy zastosowaniu kryterium MAPE_{I_E}.

Oszacowanym modelom hybrydowym nadano nazwy literowo-liczbowe. Pierwsza litera oznacza postać analityczną: liniową – L lub wykładniczą – W, a druga – kryterium wyboru najlepszego równania minimalizującego błędy prognoz: interpolacyjnych (MAPE_I) – I; ekstrapolacyjnych (MAPE_E) – E. Kolejne trzy litery – GDM, oznaczające występowanie luk w każdym z trzech rodzajów wahań sezonowych – wprowadzono ze względu na wykonane na dalszym etapie badania porównanie własności predyktywnych i prognoz otrzymanych dla najlepszych modeli z lukami występującymi w okresach godzinnych (G). Litery HG i HM odnoszą się do modeli hierarchicznych o cyklu odpowiednio dobowym i rocznym, a liczby po nich wystę-

pujące są kolejnymi podzielnikami długości cyklu wahań. W pojedynczych modelach hybrydowych występują oznaczenia 0/1, mówiące o tym, że dany rodzaj wahań jest opisywany za pomocą zmiennych zero-jedynkowych.

W dwóch ostatnich wierszach tabl. 1 i 2 podano charakterystyki liniowej i wykładniczej postaci modeli klasycznych L_GDM_0/1_0/1 i W_GDM_0/1_0/1. Dla każdej postaci analitycznej szacowano jeden model (z tego powodu po literze L lub W nie występują oznaczenia I lub E) i dlatego w przypadku kryterium $MAPE_I$ w nawiasach podano błędy prognoz ekstrapolacyjnych, a w przypadku $MAPE_E$ – błędy prognoz interpolacyjnych.

Tabl. 1. Oceny parametrów struktury stochastycznej i błędów prognoz modeli hybrydowych przy zastosowaniu kryterium $MAPE_I$

Modele	R^2	S_e w MW	V_{Se}	$MAPE_I$	$MAPE_{I_E}$
			w %		
LI_GDM_HG122_0/1	0,7980	32,76	8,969	6,665	9,817
WI_GDM_HG122_0/1	0,8106	32,49	8,823	6,584	9,307
LI_GDM_0/1_HM62	0,7827	33,99	9,306	7,123	10,480
WI_GDM_0/1_HM62	0,7982	33,55	9,108	6,985	9,947
LI_GDM_HG122_HM43	0,7710	34,87	9,547	7,330	10,620
WI_GDM_HG122_HM43	0,7860	34,53	9,375	7,248	10,110
L_GDM_0/1_0/1	0,8096	31,82	8,712	6,434	(9,076)
W_GDM_0/1_0/1	0,8227	31,45	8,540	6,308	(8,554)

Uwaga. R^2 – współczynnik determinacji, S_e – odchylenie standardowe składnika losowego, V_{Se} – współczynnik zmienności losowej.
Źródło: opracowanie własne.

Z informacji dotyczących kształtowania się wartości R^2 wynika, że wśród modeli klasycznych wyższą oceną – o 1,31 p.proc. – charakteryzuje się model wykładniczy (W_GDM_0/1_0/1), którego oceny parametrów S_e i V_{Se} mają wartości niższe odpowiednio o 0,37 MW i 0,172 p.proc. Najlepszym modelem hybrydowym okazał się jednostopniowy model wykładniczy WI_GDM_HG122_0/1 o współczynniku determinacji mniejszym o 1,21 p.proc. od lepszego z modeli klasycznych. W tym modelu wahania o cyklu dobowym zostały opisane za pomocą modelu hierarchicznego o podzielnikach 12 i 2. Oceny parametrów S_e i V_{Se} dla tego modelu przyjęły wartości nieznacznie większe od otrzymanych dla modelu klasycznego, odpowiednio o 1,04 MW i 0,28 p.proc.

Wśród modeli hybrydowych opisujących dwa rodzaje wahań lepszy okazał się model w postaci wykładniczej (WI_GDM_HG122_HM43), dla którego wartość R^2 wyniosła 0,7860. Najniższą oceną R^2 charakteryzuje się model LI_GDM_HG122 _HM43; jest ona o ok. 5,2 p.proc. niższa od otrzymanej dla modelu klasycznego w postaci wykładniczej.

Najmniejsza wartość oceny błędów $MAPE_I$ otrzymana dla modeli hybrydowych (model pojedynczy WI_GDM_HG122_0/1) jest o 0,276 p.proc. wyższa od oceny minimalnej otrzymanej dla modelu klasycznego W_GDM_0/1_0/1. Niższe oceny

błędów prognoz interpolacyjnych, podobnie jak charakterystyk własności prognostycznych, otrzymano dla predyktorów wykładniczych.

W tabl. 2 zestawiono modele hybrydowe charakteryzujące się najmniejszymi wartościami błędów $MAPE_E$ (przedostatnia kolumna). Ostatnia kolumna zawiera oceny błędów $MAPE_{E_I}$.

Tabl. 2. Oceny parametrów struktury stochastycznej i błędów prognoz modeli hybrydowych szeregu czasowego przy zastosowaniu kryterium $MAPE_E$

Modele	R^2	S_e w MW	V_{Se}	$MAPE_E$	$MAPE_{E_I}$
			w %		
LE_GDM_HG122_0/1	0,7980	32,76	8,969	9,820	6,665
WE_GDM_HG122_0/1	0,8106	32,49	8,823	9,310	6,584
LE_GDM_0/1_HM43	0,7371	37,39	10,236	9,670	7,880
WE_GDM_0/1_HM43	0,7530	37,11	10,075	9,260	7,728
LE_GDM_HG122_HM34	0,7258	38,16	10,447	9,830	8,086
WE_GDM_HG83_HM43	0,7262	39,05	10,602	9,650	8,145
L_GDM_0/1_0/1	0,8096	31,82	8,712	9,699	(6,434)
W_GDM_0/1_0/1	0,8227	31,45	8,540	9,129	(6,308)

Uwaga. Jak przy tabl. 1.
Źródło: opracowanie własne.

Najwyższą ocenę R^2 otrzymano dla pojedynczego modelu hybrydowego WE_GDM_HG122_0/1, tak samo jak w przypadku kryterium $MAPE_I$, a najniższą – dla modelu podwójnego WE_GDM_HG83_HM43. Maksymalną ocenę uzyskał ten sam model co w przypadku kryterium $MAPE_I$; jest ona niższa o 1,21 p.proc. od oceny wykładniczej postaci modelu klasycznego (W_GDM_0/1_0/1). Również dla tych samych modeli otrzymano minimalne i maksymalne oceny S_e i V_{Se}. Ocena minimalna była o ok. 0,28 p.proc. wyższa od otrzymanej dla modelu klasycznego.

Najmniejszą wartość oceny błędów $MAPE_E$ otrzymano dla predyktora opartego na modelu WE_GDM_0/1_HM43; była ona tylko o 0,053 p.proc. mniejsza od otrzymanej dla modelu o najlepszych własnościach predyktywnych WE_GDM_HG122_0/1 i zaledwie nieznacznie większa od wartości oceny błędu otrzymanego dla modelu klasycznego. Z dwóch postaci analitycznych modeli hybrydowych w przypadku tego kryterium, podobnie jak w przypadku $MAPE_I$, niższe oceny otrzymano dla postaci wykładniczej.

4. Wyniki modelowania i prognozowania dla regresyjnych modeli hybrydowych

Regresyjne równania hybrydowe, jak już wspomniano, różnią się od równań hybrydowych szeregu czasowego tym, że zawierają dodatkowo opóźnioną o 24 godziny zmienną endogeniczną oraz zmienne zero-jedynkowe oznaczające dni świąteczne i Wielką Sobotę.

W tabl. 3 zestawiono regresyjne modele hybrydowe charakteryzujące się najmniejszymi wartościami oceny błędów prognoz interpolacyjnych ($MAPE_I$). Litera P występująca na początku ich nazwy oznacza model typu regresyjnego. Własności predyktywne oraz błędy prognoz interpolacyjnych porównano z odpowiednimi charakterystykami otrzymanymi dla lepszego z predyktorów klasycznych.

Tabl. 3. Oceny parametrów struktury stochastycznej i błędów prognoz regresyjnych modeli hybrydowych przy zastosowaniu kryterium $MAPE_I$

Modele	R^2	S_e w MW	V_{Se}	$MAPE_I$	$MAPE_{I_E}$
			w %		
PLI_GDM_HG122_0/1	0,8454	28,66	7,847	5,452	7,202
PWI_GDM_HG122_0/1	0,8508	28,60	7,823	5,485	6,961
PLI_GDM_ 0/1_HM62	0,8413	29,05	7,953	5,581	7,308
PWI_GDM_ 0/1_HM62	0,8477	28,90	7,912	5,596	7,081
PLI_GDM_HG122_ HM62	0,8384	29,30	8,020	5,597	7,256
PWI_GDM_HG122_ HM62	0,8441	29,23	8,002	5,647	7,046
PL_GDM_0/1_0/1	0,8490	28,34	7,758	5,408	(7,244)
PW_GDM_0/1_0/1	0,8553	28,18	7,716	5,406	(6,983)

Uwaga. Jak przy tabl. 1.
Źródło: opracowanie własne.

Najniższa ocena R^2 charakteryzowała podwójny model liniowy PLI_GDM _HG122_HM62, a najwyższa – model pojedynczy w postaci wykładniczej PWI _GDM_HG122_0/1; była niższa tylko o 0,45 p.proc. od oceny otrzymanej dla wykładniczej postaci modelu klasycznego PW_GDM_0/1_0/1. Zwraca uwagę fakt, że rozstępy R^2 dla regresyjnych modeli hybrydowych były znacznie mniejsze od uzyskanych dla modeli szeregu czasowego i wyniosły odpowiednio 1,24 p.proc. i 3,96 p.proc.

Najniższe oceny S_e i V_{Se} otrzymano także dla modelu wykładniczego PWI_GDM_HG122_0/1. Były one nieznacznie wyższe od ocen tych parametrów charakteryzujących lepszy z modeli klasycznych (PW_GDM_0/1_0/1), odpowiednio o 0,42 MW i 0,107 p.proc. Natomiast najmniejszą wartość oceny błędu prognoz interpolacyjnych otrzymano nie dla jednostopniowego modelu wykładniczego, o najlepszych własnościach predyktywnych (PWI_GDM_HG122_0/1), lecz dla postaci liniowej tego modelu, a różnica wyniosła 0,69 p.proc. Klasyczny model wykładniczy charakteryzował się błędem mniejszym tylko o 0,10 p.proc.

Z porównania ocen błędów prognoz interpolacyjnych otrzymanych na podstawie modeli w postaci liniowej i wykładniczej o strukturze hierarchicznej tego samego rodzaju wynika, że są one, odmiennie niż w przypadku modeli szeregu czasowego, nieco wyższe dla postaci wykładniczej.

W tabl. 4 zestawiono hybrydowe modele regresyjne charakteryzujące się najmniejszymi wartościami oceny błędów $MAPE_E$. Za punkt odniesienia w porówna-

niach własności predyktywnych oraz ocen prognoz ekstrapolacyjnych przyjęto model klasyczny typu regresyjnego ze zmiennymi zero-jedynkowymi o niższej ocenie błędu tego rodzaju (PW_GDM_0/1_0/1).

Tabl. 4. Oceny parametrów struktury stochastycznej i błędów prognoz regresyjnych modeli hybrydowych przy zastosowaniu kryterium $MAPE_E$

Modele	R^2	S_e w MW	V_{Se}	$MAPE_E$	$MAPE_{E_I}$
			w %		
PLE_GDM_HG2322_0/1	0,8222	30,73	8,413	6,877	5,700
PWE_GDM_HG2322_0/1	0,8233	31,12	8,519	6,809	5,850
PLE_GDM_0/1_HM43	0,8313	29,95	8,199	6,702	5,699
PWE_GDM_0/1_HM43	0,8364	29,96	8,202	6,535	5,693
PLE_GDM_HG2322_HM43	0,8148	31,35	8,584	6,591	5,792
PWE_GDM_HG38_HM43	0,8275	30,74	8,416	6,546	5,874
PL_GDM_0/1_0/1	0,8490	28,34	7,758	7,244	(5,408)
PW_GDM_0/1_0/1	0,8553	28,18	7,772	6,983	(5,406)

Uwaga. Jak przy tabl. 1.
Źródło: opracowanie własne.

Wśród modeli klasycznych wyższą ocenę R^2 uzyskała postać wykładnicza modelu klasycznego PW_GDM_0/1_0/1; jest ona o 1,89 p.proc. wyższa od najwyższej oceny otrzymanej dla najlepszego modelu hybrydowego PWE_GDM_0/1_HM43. Najniższą ocenę S_e otrzymano dla modelu PLE_GDM_0/1_HM43, który charakteryzował się też najniższą oceną V_{Se}, a najwyższą – dla podwójnego liniowego modelu hybrydowego PLE_GDM_HG2322_ HM43.

Oceny błędów prognoz ekstrapolacyjnych ($MAPE_E$) były tylko nieznacznie zróżnicowane. Błąd prognozy dla lepszego z podwójnych modeli hybrydowych (PWE_GDM_ HG38_HM43) był większy o 0,16 p.proc. od najmniejszego błędu, a dla lepszego z modeli klasycznych (PW_GDM_0/1_0/1) – o 0,48 p.proc. większy od najmniejszego błędu otrzymanego dla regresyjnych modeli hybrydowych. Model klasyczny w rankingu 160 modeli wykładniczych (159 hybrydowych i 1 modelu klasycznego) znalazł się na 158. miejscu.

Do zbadania identyczności dystrybuant obu rozkładów empirycznych modułów względnych błędów prognoz wykorzystano składowe miernika $MAPE_E$ w liczbie 8760, oznaczone jako APE_{Et}. Posłużono się testem λ Kołmogorowa-Smirnowa (np. Domański, 1990, s. 129–130); przyjęto poziom istotności $\alpha = 0{,}05$. Statystyka empiryczna λ tego testu dla 8760 obserwacji w każdym modelu wynosi: $\lambda = D_{nm}\sqrt{n} = 0{,}03744 \cdot \sqrt{\frac{8760 \cdot 8760}{8760+8760}} = 0{,}03744 \cdot 66{,}182 = 2{,}4780$.

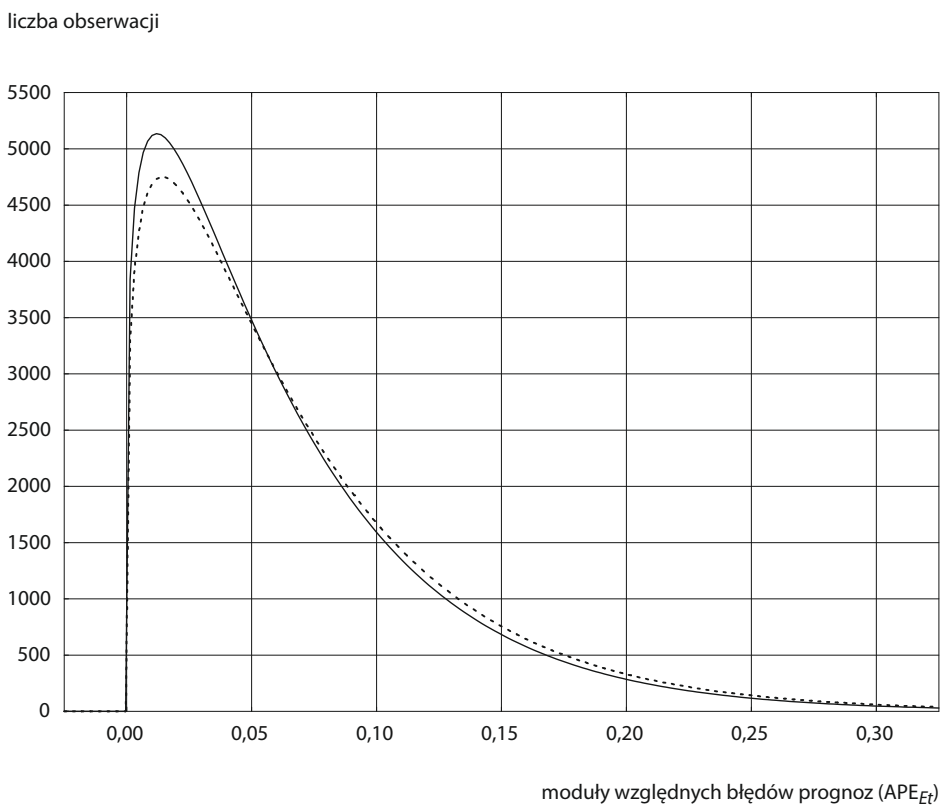
Następnie z tablicy rozkładu granicznego λ dla poziomu istotności $\alpha = 0{,}05$ odczytano wartość krytyczną $\lambda_\alpha = 1{,}354$. W teście tym obszar krytyczny jest prawo-

stronny. Ponieważ $\lambda = 2{,}4780 > 1{,}354 = \lambda_\alpha$, to hipotezę zerową należy odrzucić na korzyść hipotezy alternatywnej. Nie można więc twierdzić, że wartości APE_{Et} klasycznego modelu regresyjnego (PW_GDM_0/1_0/1) są takie same jak w przypadku regresyjnego modelu hybrydowego (PWE_GDM_0/1_HM43).

Uwzględniając powyższą decyzję weryfikacyjną oraz wcześniejsze rozważania dotyczące poziomu błędów prognoz ekstrapolacyjnych uzyskanych dla modelu klasycznego i modelu hybrydowego, można przyjąć, że błędy prognoz ekstrapolacyjnych w przypadku modelu klasycznego są znacząco większe od błędów modelu hybrydowego.

Empiryczne rozkłady modułów wartości APE_{Et} były najlepiej aproksymowane za pomocą rozkładu gamma (wykr. 2). Linią ciągłą oznaczono błędy w modelu hybrydowym PWE_GDM_0/1_HM43, a linią przerywaną – w modelu klasycznym.

Wykr. 2. Rozkład modułów względnych błędów prognoz ekstrapolacyjnych otrzymanych na podstawie modeli PWE_GDM_0/1_HM43 i PWE_GDM_0/1_0/1



Źródło: opracowanie własne.

W tabl. 5 zestawiono najlepsze modele hybrydowe szeregu czasowego i modele regresyjne charakteryzujące się najmniejszymi wartościami oceny błędów $MAPE_I$ i $MAPE_E$. Ze względu na porównywanie własności predyktywnych i prognoz otrzymanych dla modeli hybrydowych i modeli klasycznych pominięto oceny błędów $MAPE_{I_E}$ oraz $MAPE_{E_I}$. Tablica 5 zawiera także zestawienie tych wielkości dla modeli klasycznych charakteryzujących się lepszymi własnościami predyktywnymi i niższymi ocenami błędów prognoz, co pozwala na porównanie struktury modeli oraz dokładności obu rodzajów prognoz. W ostatnich dwóch wierszach podano charakterystyki własności predyktywnych i błędów prognoz otrzymane na podstawie obu rodzajów modeli klasycznych (z trzema zespołami zmiennych zero-jedynkowych) jako punkt odniesienia dla modeli hybrydowych.

Tabl. 5. Oceny parametrów struktury stochastycznej i błędów prognoz hybrydowych modeli szeregu czasowego i regresyjnych przy zastosowaniu kryteriów $MAPE_I$ oraz $MAPE_E$

Modele	R^2	S_e w MW	V_{Se}	$MAPE_I$	$MAPE_E$
			w %		
WI_GDM_HG122_0/1	0,8106	32,49	8,823	6,584	.
PLI_GDM_HG122_0/1	0,8454	28,66	7,847	5,452	.
WE_GDM_0/1_HM43	0,7530	37,11	10,075	.	9,257
PWE_GDM_0/1_HM43	0,8364	29,96	8,202	.	6,535
L_GDM_0/1_0/1	0,8227	31,45	8,540	6,308	9,129
PW_GDM_0/1_0/1	0,8553	28,18	7,772	5,406	6,983

Uwaga. Jak przy tabl. 1.
Źródło: opracowanie własne.

W pracy Szmuksta-Zawadzkiej i Zawadzkiego (2014a) przedstawiono wyniki prognozowania brakujących danych na podstawie hybrydowych modeli przyczynowo-skutkowych dla luk niesystematycznych występujących w tych samych dniach i godzinach, ale w kwietniu i październiku. Z pracy Szmuksta-Zawadzkiej i Zawadzkiego (2015, tabl. 1, s. 151) wynika, że różnica wskaźników sezonowości w przypadku tych miesięcy wynosiła 4,7 p.proc., a w przypadku sierpnia i października – 12,5 p.proc., czyli była o ok. 1,7 raza większa[4]. Pozwala to zbadać wpływ dużego wzrostu natężenia sezonowości na dokładność prognoz inter- i ekstrapolacyjnych. Natężenie sezonowości jest rozumiane jako różnica nasilenia lub skali wahań sezonowych. Zjawiskiem o większym nasileniu jest to, dla którego rozstęp wskaźników sezonowości jest większy.

[4] W tej samej tablicy, w takim samym układzie, zestawiono charakterystyki najlepszych modeli oraz błędy prognoz w wariancie występowania luk w kwietniu, a nie w sierpniu. Modele zapisano w tej samej konwencji co modele omawiane w niniejszym artykule.

Z porównania modeli hybrydowych wynika, że w przypadku obu kryteriów otrzymano modele jednostopniowe, różniące się jednak strukturą. Przy zastosowaniu kryterium $MAPE_I$ za pomocą modeli hierarchicznych WI_GDM_HG122_0/1 i PLI_GDM_HG122_0/1 opisywano wahania o cyklu dobowym o podzielnikach 12 i 2. Różniły się one postacią analityczną – w przypadku modelu szeregu czasowego była to postać wykładnicza, a w przypadku modelu regresyjnego – postać liniowa. Przy zastosowaniu kryterium $MAPE_E$ za pomocą wykładniczych modeli hierarchicznych WE_GDM_ 0/1_HM43 i PWE_GDM_0/1_HM43 opisywano wahania o cyklu 12-miesięcznym.

Z porównania własności predyktywnych hybrydowych równań szeregu czasowego i równań regresyjnych wynika, że lepszymi własnościami charakteryzują się modele regresyjne. W przypadku kryteriów $MAPE_I$ i $MAPE_E$ współczynnik determinacji przyjął wartości większe odpowiednio o 4,47 i 8,11 p.proc. Modele regresyjne charakteryzowały się także mniejszymi wartościami odchylenia standardowego składników losowych i współczynnika zmienności losowej. Różnice te wynosiły odpowiednio 3,83 MW i 7,15 MW oraz 0,976 p.proc. i 1,273 p.proc. Podane liczby jednoznacznie wskazują na lepsze własności predyktywne modeli regresyjnych. Przy zastosowaniu obu kryteriów otrzymano także niższe oceny błędów prognoz. W przypadku prognoz interpolacyjnych różnica wyniosła 1,132 p.proc. (17,2%), a w przypadku prognoz ekstrapolacyjnych – 2,722 p.proc. (29,4%).

W wykładniczych modelach klasycznych zarówno własności predyktywne, jak i błędy prognoz przyjęły korzystniejsze wartości także dla predyktora regresyjnego. Współczynnik determinacji w modelach hybrydowych modeli regresyjnych osiągnął wartości nieznacznie mniejsze niż w modelu klasycznym; dla kryteriów $MAPE_I$ i $MAPE_E$ różnice te wyniosły odpowiednio 0,99 p.proc. i 1,89 p.proc., a w przypadku odchylenia standardowego składników losowych i współczynnika zmienności losowej – odpowiednio 0,48 MW i 1,78 MW oraz 0,075 p.proc. i 0,430 p.proc. Błąd prognoz interpolacyjnych dla regresyjnego modelu hybrydowego wyniósł 5,452% i był większy o 0,046 p.proc. od błędu dla modelu klasycznego.

Z porównania własności predyktywnych i dokładności prognoz dotyczących modeli różniących się natężeniem sezonowości w przypadku miesięcy, w których wystąpiły luki w danych, wynika, że wartości współczynnika determinacji, parametrów struktury stochastycznej i przeciętnego błędu względnego prognoz są zbliżone. Istotny okazał się natomiast wpływ różnicy ocen wskaźników sezonowości na dokładność prognoz ekstrapolacyjnych. W przypadku modelu o mniejszej różnicy wartości tych wskaźników otrzymano błędy mniejsze o ok. 1,7 p.proc. (35%), a w przypadku modeli klasycznych różnica ta wyniosła ok. 1,4 p.proc. (26%).

5. Porównanie wyników modelowania i prognozowania dla wariantów luk G i GDM

Porównanie dokładności opisu modeli i błędów prognoz otrzymanych dla najlepszych modeli z lukami występującymi w okresach godzinnych (G) i rozpatrywanymi wcześniej modelami z lukami występującymi w każdym z trzech rodzajów wahań (GDM) pozwala na dokonanie oceny wpływu zwiększenia liczby luk – w omawianym badaniu o ok. 80% – na dokładność opisu oraz prognoz.

Analizę przeprowadzono osobno dla modeli szeregu czasowego i modeli regresyjnych; uwzględniono także modele klasyczne. Ze względu na porównywanie własności predyktywnych i prognoz otrzymanych dla modeli hybrydowych i modeli klasycznych pominięto (podobnie jak w tabl. 5) oceny błędów $MAPE_{I_E}$ oraz $MAPE_{E_I}$. Informacje dotyczące wariantu luk G zostały zaczerpnięte z pracy Zawadzkiego (2018).

Tabl. 6. Porównanie modeli szeregu czasowego w wariantach luk G i GDM przy zastosowaniu kryteriów $MAPE_I$ oraz $MAPE_E$

Modele	R^2	S_e w MW	V_{Se}	$MAPE_I$	$MAPE_E$
			w %		
WI_G _HG432_0/1	0,7661	36,04	9,785	6,391	.
WI_GDM_HG122_0/1	0,8106	32,49	8,823	6,584	.
WE_G_ HG122_0/1	0,8106	32,44	8,808	.	8,775
WE_GDM_0/1_HM43	0,7530	37,11	10,075	.	9,257
W_G_0/1_0/1	0,8207	31,57	8,572	6,194	8,554
W_GDM_0/1_0/1	0,8227	31,45	8,540	6,308	9,129

Uwaga. Jak przy tabl. 1.
Źródło: opracowanie własne.

Przy zastosowaniu kryterium $MAPE_I$ model hybrydowy z mniejszą liczbą luk (WI_G_HG432_0/1) otrzymał niższą o ponad 4,45 p.proc. ocenę R^2 oraz wyższe oceny S_e i V_{Se}, odpowiednio o 3,55 MW i 0,962 p.proc., ale niższą o ok. 0,2 p.proc. ocenę błędu prognoz interpolacyjnych niż model WI_GDM_HG122_0/1. Z kolei przy zastosowaniu kryterium $MAPE_E$ model WE_G_ HG122_0/1 charakteryzował się zarówno korzystniejszymi własnościami predyktywnymi, jak i niższymi wartościami błędu prognoz. Wartość współczynnika R^2 była o 5,76 p.proc. mniejsza od otrzymanej dla modelu WE_GDM_0/1_HM43, a wartości współczynników S_e i V_{Se} były mniejsze odpowiednio o 4,67 MW i 1,266 p.proc. Dla modelu o mniejszej liczbie luk otrzymano także niższą – o 0,482 p.proc. (5,5%) – ocenę błędu prognoz ekstrapolacyjnych. Różnicę tę w przypadku liczby luk większej o ponad 80% należy uznać za niewielką. W prognozach interpolacyjnych wynosiła ona 1,03%.

Z powyższych rozważań można wyprowadzić wniosek, że dla celów prognozowania pod uwagę należy brać dwa kryteria wyboru równań: $MAPE_I$ i $MAPE_E$.

Z informacji zawartych w tabl. 6 wynika ponadto, że modele klasyczne, podobnie jak wszystkie modele hybrydowe, są modelami wykładniczymi. Nieznacznie lepszymi własnościami predyktywnymi charakteryzował się model W_GDM_0/1_0/1, ale niższe oceny błędów obu rodzajów prognoz otrzymano dla modelu W_G_0/1_0/1. W przypadku prognoz interpolacyjnych błąd był mniejszy o 1,08 p.proc., a w przypadku prognoz ekstrapolacyjnych – o 6,72 p.proc. Oznacza to, że podobnie jak w modelach hybrydowych, wzrost liczby luk w prognozach ekstrapolacyjnych był nieco silniejszy. Oceny błędu prognoz interpolacyjnych otrzymane dla modeli klasycznych były niższe nie więcej niż o 0,40 p.proc., a prognoz ekstrapolacyjnych – co najwyżej o 0,70 p.proc.

W tabl. 7 zestawiono parametry charakteryzujące własności predyktywne i błędy prognoz najlepszych modeli regresyjnych z lukami występującymi w okresach godzinnych (G) i modeli z lukami w każdym z trzech rodzajów wahań (GDM).

Tabl. 7. Porównanie hybrydowych modeli regresyjnych dla wariantów luk G i GDM przy zastosowaniu kryteriów $MAPE_I$ i $MAPE_E$

Modele	R^2	S_e w MW	V_{Se}	$MAPE_I$	$MAPE_E$
			w %		
PLI_G_HG432_0/1	0,8344	29,63	8,107	5,406	.
PLI_GDM_HG122_0/1	0,8454	28,66	7,847	5,452	.
PWE_G_HG432_0/1	0,8379	29,75	8,145	.	6,622
PWE_GDM_0/1_HM43	0,8364	29,96	8,202	.	6,535
PL_G_0/1_0/1	0,8485	28,36	7,76	5,374	6,906
PW_GDM_0/1_0/1	0,8553	28,18	7,716	5,406	6,983

Uwaga. Jak przy tabl. 1.
Źródło: opracowanie własne.

Z informacji zawartych w tabl. 7 wynika, że przy zastosowaniu kryterium $MAPE_I$ najlepszymi modelami regresyjnymi okazały się jednostopniowe hybrydowe modele liniowe z różnymi podzielnikami dla 24-godzinnego cyklu wahań. W wariancie G był to hierarchiczny model trójstopniowy z podzielnikami wynoszącymi odpowiednio 4, 3 i 2. W wariancie ze znacznie większą liczbą luk (GDM) był to model dwustopniowy o podzielnikach 12 i 2. Przy zastosowaniu kryterium $MAPE_E$ oba predyktory przyjęły postać wykładniczą, przy czym w przypadku mniejszej liczby luk był to model o tych samych podzielnikach co przy zastosowaniu kryterium $MAPE_I$, ale

w postaci wykładniczej, a w wariancie GDM – model dwustopniowy z podzielnikami cyklu rocznego wynoszącymi 4 i 3.

W modelach klasycznych otrzymano predyktory różniące się postacią analityczną – w wariancie G lepsza okazała się postać liniowa, a w wariancie GDM – wykładnicza.

Porównanie dokładności opisu i błędów prognoz najlepszych modeli regresyjnych, zarówno hybrydowych, jak i klasycznych, w obu wariantach luk pokazuje, że są one bardzo zbliżone. Różnica wartości R^2 między wartością maksymalną, otrzymaną dla modelu klasycznego PW_GDM_0/1_0/1, a wartością minimalną, otrzymaną dla hybrydowego modelu liniowego PLI_G_HG432_0/1, wyniosła 2,1 p.proc, a w przypadku modeli hybrydowych przy zastosowaniu kryterium $MAPE_I$ – 1,1 p.proc.

Dla modeli klasycznych różnica ocen R^2 wynosi 0,68 p.proc., przy czym większą wartością charakteryzuje się wariant GDM. Podobnie wygląda sytuacja w przypadku modeli hybrydowych, przy zastosowaniu kryterium $MAPE_I$. Relacje te potwierdzają niższe oceny S_e i V_{Se}.

Przy zastosowaniu kryterium $MAPE_E$ różnice ocen parametrów struktury stochastycznej modeli hybrydowych w wariantach G i GDM są niewielkie. W przypadku R^2 niższą – o 0,15 p.proc. – ocenę otrzymano dla wariantu G. Ocena S_e dla modelu z mniejszą liczbą luk jest niższa o 0,21 MW, a ocena V_{Se} – o 0,57 p.proc.

Różnica ocen parametrów R^2, S_e i V_{Se} między lepszym z modeli klasycznych (PW_GDM_0/1_0/1) a modelem PWE_G_HG432_0/1 wynosi odpowiednio 1,06 p.proc., 1,39 MW oraz 0,429 p.proc.

Prawidłowości dotyczące zbliżonych własności predyktywnych znajdują pełne odzwierciedlenie w dokładności prognoz. Przy zastosowaniu kryterium $MAPE_I$ model o mniejszej liczbie luk uzyskał ocenę niższą o 0,163 p.proc.

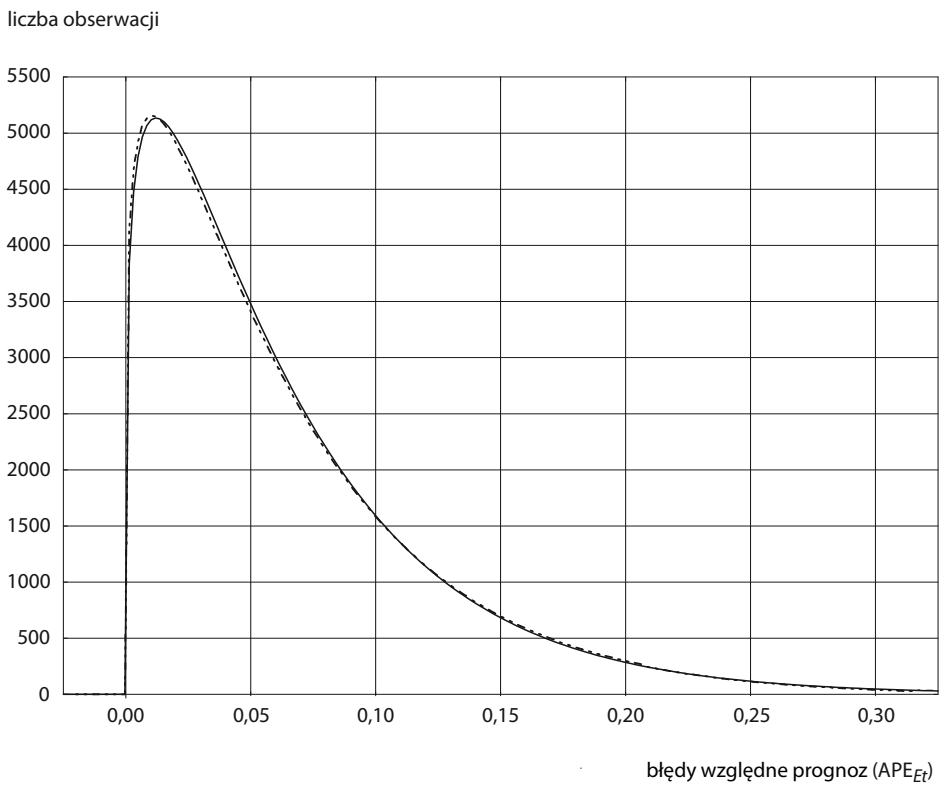
Bardzo zbliżone są także oceny błędów $MAPE_E$ otrzymanych na podstawie najlepszych modeli regresyjnych w wariantach G (PWE_G_HG432_0/1) i GDM (PWE_GDM_ 0/1_HM43). W wariancie GDM (z liczbą luk większą o ok. 80%) błąd ten był nawet o 0,087 p.proc. (1,33%) mniejszy niż w wariancie G.

Do zbadania identyczności dystrybuant obu rozkładów empirycznych modułów względnych błędów APE_{Et} wykorzystano, podobnie jak poprzednio, test λ Kołmogorowa-Smirnowa, przy przyjęciu tego samego poziomu istotności (α = 0,05). Statystyka empiryczna λ tego testu, dla 8760 obserwacji w każdym modelu, wynosi $\lambda = D_{nm}\sqrt{n} = 0{,}006164 \cdot 66{,}182 = 0{,}4080$. Z tablicy rozkładu granicznego λ odczytujemy wartość krytyczną $\lambda_{0,05} = 1{,}354$. Ponieważ $\lambda = 0{,}4080 < 1{,}354 =$

$= \lambda_{0,05}$, to nie ma podstaw do odrzucenia hipotezy zerowej, w której założono identyczność badanych rozkładów.

Wykres 3 przedstawia aproksymanty teoretyczne rozkładów empirycznych modułów APE_{Et}. Najlepiej były one aproksymowane, podobnie jak poprzednio, za pomocą rozkładu gamma.

Wykr. 3. Aproksymanty teoretyczne rozkładu empirycznego modułów APE_{Et} otrzymanych na podstawie modeli PWE_GDM_0/1_HM43 i PWE_G_HG432_0/1



Źródło: opracowanie własne.

Z porównania rozkładów błędów prognoz ekstrapolacyjnych w wariantach G i GDM wynika, że są one bardzo zbliżone.

Ogólny wniosek z powyższych porównań jest taki, że modele w wariancie GDM, o liczbie luk większej o 80% niż w wariancie G, charakteryzowały się zarówno bardzo zbliżonymi własnościami predyktywnymi, jak i ocenami błędów prognoz inter- i ekstrapolacyjnych.

6. Wnioski

Z analiz przeprowadzonych na omawianym szeregu czasowym można wyprowadzić kilka istotnych wniosków szczegółowych.

W przypadku podwójnych modeli hybrydowych szeregu czasowego i modeli regresyjnych podstawą wyboru powinny być kryteria dotyczące minimalizacji błędów prognoz interpolacyjnych ($MAPE_I$) i ekstrapolacyjnych ($MAPE_E$).

Regresyjne modele hybrydowe charakteryzowały się lepszymi własnościami prognostycznymi i niższymi ocenami błędów prognoz inter- i ekstrapolacyjnych niż modele szeregu czasowego, przy czym różnice w dokładności były większe w przypadku prognoz ekstrapolacyjnych.

We wszystkich przypadkach najlepszymi modelami hybrydowymi okazały się pojedyncze modele hierarchiczne opisujące wahania o cyklu 24-godzinnym. Otrzymano dla nich mniejsze błędy prognoz niż dla modeli pojedynczych opisujących wahania o cyklu 12-miesięcznym i dla modeli podwójnych.

Własności predyktywne i błędy prognoz interpolacyjnych otrzymanych na podstawie obu klas modeli oraz prognoz ekstrapolacyjnych w wariancie, w którym luki występowały tylko w ośmiu okresach godzinnych (G), były tylko nieznacznie mniejsze niż w wariancie z lukami w ośmiu okresach godzinnych, dwóch dniach tygodnia i dwóch miesiącach (GDM), mimo że w drugim wariancie liczba luk była większa o ponad 80%.

Ocena przeciętnego względnego błędu prognoz ekstrapolacyjnych otrzymana dla modelu regresyjnego w wariancie GDM była o ponad 5 p.proc. niższa od otrzymanej dla modelu klasycznego w wariancie G. Oceny błędów prognoz ekstrapolacyjnych otrzymane dla modeli regresyjnych w obu wariantach luk były zbliżone i jednocześnie o kilka punktów procentowych niższe od oceny błędów otrzymanych dla modeli klasycznych. Niższe oceny błędów prognoz ekstrapolacyjnych przemawiają za wykorzystaniem modeli hybrydowych w prognozowaniu zmiennych z potrójnie złożonymi wahaniami sezonowymi.

Przeprowadzone badanie prowadzi do konkluzji, że modele hybrydowe szeregu czasowego i hybrydowe modele regresyjne mogą być użytecznym narzędziem modelowania i prognozowania w warunkach braku pełnej informacji. Mimo że wariant GDM charakteryzował się większą aż o 80% liczbą luk niż wariant G, to zarówno własności predyktywne, jak i oceny błędów prognoz inter- i ekstrapolacyjnych były bardzo zbliżone.

Bibliografia

Allisson, P. D. (2001). *Missing Data*. Thousand Oaks: Sage Publications.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., Petropoulos, F. (2017). Forecasting with Temporal Hierarchies. *European Journal of Operational Research*, *262*(1), 60–74. DOI: 10.1016/j.ejor.20 17.02.046.

Domański, Cz. (1990). *Testy statystyczne*. Warszawa: Państwowe Wydawnictwo Ekonomiczne.

Henn, B., Raleigh, M. S., Fisher, A., Lundquist, J. D. (2013). A Comparision of Methods for Filling Gaps in Hourly Near-Surface Air Temperature Data. *Journal of Hydrometorology*, *14*(3), 929–945. DOI: 10.1175/JHM-D-12-027.1.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics* & *Data Analysis*, *55*(9), 2579–2589. DOI: 10.1016/j.csda.2011.03.006.

Kufel, T. (2010). *Ekonometryczna analiza cykliczności procesów gospodarczych o wysokiej częstotliwości obserwwowania*. Toruń: Wydawnictwo Naukowe UMK.

Kwac, J., Flora, J., Rajagopal, R. (2014). Household Energy Consumption Segmentation Using Hourly Data. *IEEE Transactions on Smart Grid, 5*(1), 420–430. DOI: 10.1109/TSG.2013.2278477.

Little, R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data: Second Edition*. Hoboken, New Jersey: John Wiley & Sons. DOI: 10.1002/9781119013563.

Raudenbush, S. W., Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, London, New Delhi: Sage Publication Inc. Pobrane z: https://books.google.pl/books?id=uyCV0CNGDLQC&pg=PR3&hl=pl&source=gbs_selected_pages&cad=2#v=onepage&q&f=false.

Szmuksta-Zawadzka, M., Zawadzki, J. (2002). Hierarchiczne modele szeregów czasowych z wahaniami sezonowymi. Budowa. Estymacja. Prognozowanie. W: A. Zeliaś (red.), *Przestrzenno--czasowe modelowanie i prognozowanie zjawisk gospodarczych* (s. 193–204). Kraków: Wydawnictwo AE.

Szmuksta-Zawadzka, M., Zawadzki, J. (2011). Modelowanie i prognozowanie zapotrzebowania na energię elektryczną w mikroskali. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, (165), 152–161.

Szmuksta-Zawadzka, M., Zawadzki, J. (2014a). Modele hierarchiczne w prognozowaniu zmiennych o wysokiej częstotliwości obserwowania w warunkach braku pełnej informacji. *Ekonometria. Econometrics*, *46*(4), 72–84. DOI: 10.15611/ekt.2014.4.07.

Szmuksta-Zawadzka, M., Zawadzki, J. (2014b). Wykorzystanie modeli „hybrydowych" w prognozowaniu brakujących danych w szeregach ze złożoną okresowością (sezonowością). W: W. Jurek (red.), *Matematyka i informatyka na usługach ekonomii*: *rozważania ogólne, modele* (s. 72–83). Poznań: Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu.

Szmuksta-Zawadzka, M., Zawadzki, J. (2015). Wykorzystanie danych oczyszczonych o wysokiej częstotliwości w prognozowaniu zmiennych ze złożoną sezonowością. *Metody Ilościowe w Badaniach Ekonomicznych*, *16*(4), 147–159. Pobrane z: http://qme.sggw.pl/pdf/MIBE_T16_z4.pdf.

Taieb, S. B., Yu, J., Barreto, M. N., Rajagopal, R. (2017). *Regularization in Hierarchical Time Series Forecasting with Application to Electricity Smart Meter Data*. Pobrane z: https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14173/14055.

Weron, R. (2006). *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. Chichester: John Wiley & Sons.

Witkowska, D., Górecka, A., Szadkowska, D., Szymczak, Z. (2000). The forecasting of the demand for electric energy: comparative analysis. *Dynamic Econometric Models*, *4*, 45–59.

Zawadzki, J. (2018). Modele hybrydowe w prognozowaniu brakujących danych w szeregach o bardzo wysokiej częstotliwości obserwowania. *Folia Pomeranae Universitatis Technologiae Stetinensis*, *Oeconomica*, *346*(92), 81–96. Pobrane z: http://foliaoe.zut.edu.pl/pdf/346/oe2018.92.3.07.pdf.