EKONOMETRIA ECONOMETRICS 4(38) • 2012

ISSN 1507-3866

Maria Szmuksta-Zawadzka, Jan Zawadzki
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

O METODZIE PROGNOZOWANIA BRAKUJĄCYCH DANYCH W SZEREGACH CZASOWYCH O WYSOKIEJ CZĘSTOTLIWOŚCI Z LUKAMI SYSTEMATYCZNYMI

Streszczenie: W pracy podjęta zostanie próba rozszerzenia rozważań zawartych w artykule [Szmuksta-Zawadzka, Zawadzki 2011] dotyczących prognozowania w szeregach czasowych o wysokiej częstotliwości obserwowania z lukami niesystematycznymi na przypadek występowania luk systematycznych. Rozważania teoretyczne zostaną zilustrowane przykładem empirycznym dotyczącym modelowania i prognozowania zapotrzebowania na moc energetyczną w okresach półgodzinnych w jednej z aglomeracji miejskich.

Słowa kluczowa: szeregi czasowe o wysokiej częstotliwości obserwowania, prognozowanie brakujących danych, luki systematyczne.

1. Wstęp

W pracy [Szmuksta-Zawadzka, Zawadzki 2011] przedstawiono wyniki modelowania i prognozowania zapotrzebowania na moc energetyczną dla niesystematycznych luk w danych o wysokiej częstotliwości. Wykazano w niej przydatność modeli szeregu czasowego z trzema zespołami zmiennych zero-jedynkowych opisujących wahania o cyklach: rocznym, tygodniowym i dobowym obejmującym 48 okresów półgodzinnych w prognozowaniu brakujących danych dla tego rodzaju luk. W niniejszej pracy podjęta zostanie próba uogólnienia rozważań na przypadek występowania luk systematycznych, tzn. taki, gdy w przedziale czasowym „próby" (okresie estymacyjnym) nie są dostępne dane statystyczne przynajmniej o jednym podokresie wchodzącym w skład określonego cyklu.

Praca składa się z dwóch części. W części pierwszej o charakterze teoretycznym przedstawione zostaną rozważania dotyczące prognozowania brakujących danych w sytuacji, gdy w szeregach czasowych o wysokiej częstotliwości występują luki systematyczne. W części drugiej przedstawione zostaną wyniki prognozowania inter- i ekstrapolacyjnego zapotrzebowania na moc energetyczną w okresach półgodzinnych w jednej z aglomeracji miejskich.

2. Metodologia prognozowania brakujących danych w szeregach czasowych

Do opisu wahań okresowych w szeregach czasowych wykorzystuje się bądź zmienne zero-jedynkowe przyjmujące wartość 1 w okresie k i zero w pozostałych m-1 okresach cyklu, bądź wielomiany trygonometryczne o składowych sinusoidalnych i kosinosuidalnych.

Z literatury przedmiotu wiadomo, że u podstaw modelowania i prognozowania zmiennych z wahaniami sezonowymi leży koncepcja modelowania oszczędnego (por. np. [Dittmann 2000; Szmuksta-Zawadzka, Zawadzki 1998; 2000; 2006; Zeliaś i in. 2003]). Polega ona na uwzględnieniu w modelu tylko statystycznie istotnych składowych opisujących wahania tego rodzaju.

Sposób opisu ma istotne znaczenie dla przebiegu procesu modelowania i prognozowania brakujących danych, zwłaszcza w szeregach z lukami systematycznymi.

W przypadku występowania luk niesystematycznych ma to znaczenie tylko z punktu widzenia liczby stopni swobody. Jest to jednak szczególnie istotne, gdy w modelowaniu wykorzystuje się krótkie szeregi czasowe o okresie jednostkowym wynoszącym jeden miesiąc lub jedną dekadę.

Występowanie systematycznych luk w danych pociąga za sobą następstwa idące znacznie dalej, i to niezależnie od długości cyklu wahań.

W modelach ze zmiennymi zero-jedynkowymi nieistotne z definicji są składowe odpowiadające podokresom cyklu, w których występują luki systematyczne. Dzieje się tak dlatego, że przyjmują one wartości zerowe i tym samym mają stałe wariancje.

W modelach z wielomianem trygonometrycznym dla pełnych danych poszczególne składowe są ze sobą nieskorelowane. Fakt ten umożliwia badanie udziałów zarówno poszczególnych składowych, jak i harmonik w wyjaśnianiu wariancji sezonowej zmiennej prognozowanej. Występowanie luk systematycznych sprawia, że nie można obliczyć tych udziałów, ponieważ w zależności od liczby i układu luk część składowych może być dokładnie współliniowa (tzn. współczynniki korelacji liniowej będą przyjmować wartości plus lub minus jeden), a większość składowych harmonicznych będzie ze sobą skorelowana. Ponadto, niektóre ze składowych mogą przyjmować stałe wartości, a więc zostaną one automatycznie wyłączone z procesu modelowania. Ponadto część składowych może być kombinacjami liniowymi innych składowych.

Wiązać się to będzie z koniecznością szacowania wielu wersji modeli zawierających składowe nietworzące kombinacji liniowych. Wersje te są nierozróżnialne z punktu widzenia własności predyktywnych, tzn. mają identyczne oceny współczynników determinacji, odchyleń standardowych składników losowych oraz statystyk DW. Natomiast będą się one różnić dokładnością prognoz inter- i ekstrapolacyjnych.

W dalszej części artykułu zajmować się będziemy modelami ze zmiennymi zero-jedynkowymi.

W modelach tych, ze względów estymacyjnych, szacuje się o jeden parametr mniej niż wynosi długość cyklu wahań (m). Wyraz wolny interpretuje się w różny sposób w zależności od tego, czy w jednym z wierszy (odpowiadającym zazwyczaj ostatniemu okresowi cyklu) występują elementy równe minus jeden czy zero. W pierwszym przypadku będzie on średnią arytmetyczną obliczoną ze stałych parametrów trendów liniowych dla wszystkich podokresów. Natomiast ocena odchylenia sezonowego dla m-tego podokresu będzie wziętą z przeciwnym znakiem sumą odchyleń z pozostałych m-1 podokresów. Wynika to z nałożonego warunku sumowalności do zera odchyleń sezonowych. Występowanie elementów równych zeru oznacza, że wyraz wolny równy jest wyrazowi wolnemu trendu dla m-tego podokresu. Zatem odchylenia sezonowe dla pozostałych podokresów będą odchyleniami od tak zdefiniowanego wyrazu wolnego. Fakt ten nie ma większego znaczenia w sytuacji, gdy parametry modelu szacowane są na podstawie szeregu niezawierającego luk systematycznych. Oznacza to jedynie przeskalowanie parametrów. Posiada on istotne znaczenie w przypadku występowania luk systematycznych, ponieważ występowanie elementów równych zeru lub równych minus jeden będzie wpływać na wielkości prognoz.

Istnieje zatem możliwość „manipulowania" wyrazem wolnym w taki sposób, aby punktem odniesienia był podokres najbardziej „podobny" do podokresów, w których występują luki systematyczne.

Skorzystanie z tego sposobu możliwe jest jedynie wtedy, gdy luki występować będą w jednym albo co najwyżej w dwóch podokresach o zbliżonych odchyleniach sezonowych. Jeżeli luk jest więcej i odnoszą się one do podokresów cykli o różnej długości, jak ma to miejsce w szeregach czasowych o wysokiej częstotliwości obserwowania, „manipulowanie" wyrazem wolnym nie wchodzi w grę.

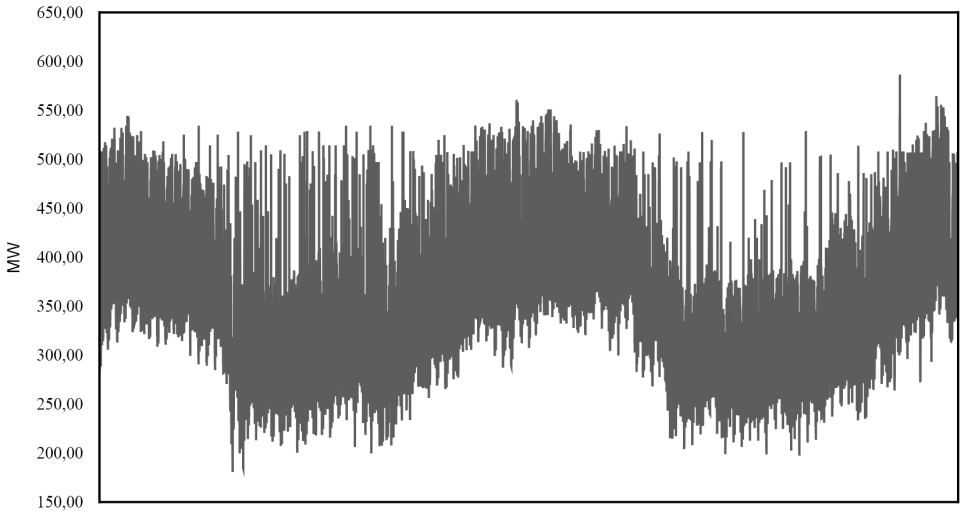
Można natomiast zaproponować zastosowanie sposobu polegającego na tym, że oceny odchyleń sezonowych dla podokresów z lukami będą średnimi z odchyleń okresów poprzedzających i po nich następujących. Tym średnim będą równe oczywiście także odchylenia sezonowe dla podokresów, które posłużyły do ich wyznaczenia. Za wykorzystaniem tej metody przemawia fakt, że dla okresów sąsiadujących ze sobą oceny parametrów będą zazwyczaj mniej się różnić od oceny dla mniej lub bardziej arbitralnie wybranego podokresu. Ponadto można oczekiwać, że średnie będą „niwelować" występowanie większych zakłóceń o charakterze losowym.

3. Prognozy brakujących danych na przykładzie zapotrzebowania na moc energetyczną

Modelowanie i prognozowanie brakujących danych półgodzinnych w szeregu czasowym dotyczyć będzie zapotrzebowania na moc energetyczną w jednej z aglomeracji dla wybranego wariantu luk systematycznych.

Kształtowanie się zmiennej w okresie estymacyjnym obejmującym dwa lata (35 040 obserwacji) zostało przedstawione na rys. 1. Trzeci rok będzie okresem em-

pirycznej weryfikacji prognoz. Dla tego roku zbudowane zostaną prognozy ekstrapolacyjne.



Rys. 1. Kształtowanie się zapotrzebowania na energię elektryczną w okresach półgodzinnych (w MW)

Źródło: [Szmuksta-Zawadzka, Zawadzki 2011].

W modelowaniu zostanie wykorzystany, podobnie jak dla danych z lukami niesystematycznymi, model zawierający trzy zespoły zmiennych zero-jedynkowych opisujące wahania o cyklach: rocznym (12-miesięcznym), tygodniowym (7-dniowym) i dobowym obejmującym 48 okresów półgodzinnych. Szacowane więc będą parametry równania:

$$Y_t = \alpha_1 t + \alpha_{12,7,48} + \sum_{i=1}^{11} b_{0i} M_{it} + \sum_{j=1}^{6} c_{0j} D_{jt} + \sum_{k=1}^{47} d_{ok} P_{kt} + \sum_{l=1}^{r} a_{l0} S_{lt} + \delta Y_{t-48} + U_t, \qquad (1)$$

gdzie: M_{it} – miesiąc,

D_{jt} – dzień tygodnia,

P_{kt} – okres półgodzinny w cyklu dziennym.

W modelu (1) występować będą także zmienne oznaczające występowanie świąt oraz Wielkiej Soboty (S_{lt}) oraz opóźniona o jedną dobę zmienna prognozowana ($Y_{t\text{-}48}$).

Z zapisu wyrazu wolnego w powyższym równaniu wynika, że w macierzy X w wierszach dla zmiennych zero-jedynkowych odpowiadających ostatnim okresom cykli występowały elementy równe zeru. Zatem wyraz wolny $\alpha_{12,7,8}$ jest wielkością

odnoszącą się do grudnia, siódmego dnia tygodnia oraz 48 okresu półgodzinnego (godz. 0^{00}). W naszym przypadku siódmym dniem tygodnia był piątek, ponieważ w sobotę przypadał Nowy Rok, będący pierwszym dniem, z którego pochodziły dane. Oznacza to tym samym, że parametry: b_{0i}, c_{0j}, d_{0k} interpretowane będą jako odchylenia od odpowiedniego elementu składowego wyrazu wolnego.

Rozpatrywany będzie jeden wariant luk systematycznych, w którym założono, że braki danych występują w składowych odpowiadających każdemu rodzajowi wahań. Luki systematyczne występować będą:

– w marcu,
– we wtorki,
– w: 2, 8, 14, 20, 26, 32, 38 i 44 okresie półgodzinnym.

Oznacza to, że luki będą występować łącznie w 12 160 okresach spośród 35 020 okresów półgodzinnych. Stanowić zatem będą ok. 35% długości szeregu czasowego.

Szacowane będą modele ze zmiennymi zero-jedynkowymi dla pełnych danych oraz dla danych z lukami – oznaczono je odpowiednio przez: PEL_01 oraz LUK_01. Estymacji poddano także modele, w których parametry dla podokresów z lukami są średnimi z podokresów poprzedzających i następujących po nich – oznaczono je jako PEL_SR oraz LUK_SR. Zawierać one będą zmienne dla podokresów, które zostały wzięte do obliczeń średnich.

W tabeli 1 zestawione zostały oszacowania parametrów strukturalnych wymienionych wyżej modeli. W ostatnich czterech wierszach podane zostały syntetyczne oceny parametrów struktury stochastycznej. Prezentując wyniki, symbole zmiennych M_{it} oraz D_{jt} odpowiadające miesiącom w roku oraz dniom tygodnia zastąpiono skrótami ich nazw. Było to uzasadnione zwłaszcza w przypadku dni, ponieważ siódmym dniem w cyklu tygodniowym był piątek. Podobnie postąpiono w przypadku świąt, podając skrótową ich nazwę lub skrótowa datę, kiedy one przypadają.

Występowanie pustych komórek oznacza, że dana zmienna nie wystąpiła w modelu. Z kształtowania się współczynników determinacji wynika, że są bardzo zbliżone. Ocenami niższymi o ok. 1,5 punktu-2 punkty procentowe charakteryzują się modele z uśrednionymi parametrami (SR). Jednocześnie charakteryzują się one wyższymi o ok. 1,5-2 MW ocenami odchyleń standardowych składników losowych i wyższymi co najwyżej o ok. 0,7 punktu procentowego ocenami współczynników zmienności losowej.

Nieznacznie gorsze własności predyktywne równań SR wynikają ze zmniejszonej liczby szacowanych parametrów. Porównanie stopnia opisu modeli dla pełnych danych (PEL) z modelami, których parametry szacowane były na podstawie szeregów zawierających luki systematyczne (LUK), wypada także nieznacznie lepiej dla modeli pełnych. Różnice w ocenach współczynników determinacji wynoszą ok. 2 punkty procentowe, a współczynników zmienności losowej 0,3-0,4 punktu procentowego.

Podsumowując krótko analizę kształtowania się ocen parametrów struktury stochastycznej, możemy powiedzieć, że zarówno rodzaj modelu, jak i liczba obserwa-

cji nie miały większego wpływu na własności predyktywne równań. Z porównania ocen wyrazów wolnych wynika, że w modelu PEL_SR jest ona o ok. 10 MW wyższa niż w pozostałych modelach. Natomiast oceny parametrów przy tych samych zmiennych odpowiadających poszczególnym podokresom różnią się najwyżej o ok. 4 MW, tj. tylko nieco więcej niż 1%. Oceny parametrów przy zmiennej *t* są bliskie zeru, jedynie w modelu ze zmiennymi zero-jedynkowymi dla pełnych danych ocena parametru jest dodatnia, w pozostałych trzech ujemna. Największe różnice ocen parametrów dla dni świątecznych między modelami ze zmiennymi zero-jedynkowymi i z uśrednionymi parametrami występują dla Wielkiej Soboty i Wielkanocy oraz Nowego Roku. Wynoszą one odpowiednio ok. 20 i 12 MW. Dla pozostałych dni świątecznych i zmiennych grupujących w modelach SR są one bardzo zbliżone. Bardzo zbliżone są także oceny przy zmiennej $Y_{t\text{-}48}$ – występujące przy niej współczynniki „przeniesienia" kształtują się na poziomie 43-44%. Natomiast różnią się, i to niekiedy znacznie, zarówno co do znaku, jak i co do wielkości, oceny parametrów w ramach poszczególnych rodzajów wahań.

W przypadku miesięcy amplituda między styczniem a majem przekracza 70 MW. Dla cyklu tygodniowego różnica ocen między poniedziałkiem a niedzielą wynosi ok. 60 MW. Dla cyklu dobowego różnica ocen parametrów przy zmiennych P9 (godz. 4^{30}) a P41 (godz. 20^{30}) przekracza 60 MW. Występują znaczne różnice także pomiędzy świętami. Amplituda ocen między Świętami Bożego Narodzenia a dniem Wszystkich Świętych przekracza 80 MW.

W tabeli 2 zestawione zostały oceny błędów prognoz interpolacyjnych. Spośród 12 160 prognoz interpolacyjnych tylko 2976 (31 dni * 48 okresów półgodzinnych * 2 lata) odnosi się łącznie do marca i obejmuje wszystkie dni tygodnia i wszystkie okresy półgodzinne. Dla pozostałych 11 miesięcy prognozy w liczbie 9184 obejmują w całości tylko wtorki oraz 8 okresów półgodzinnych dla dni tygodnia oprócz wtorku. To samo odnosi się do dni świątecznych nieprzypadających we wtorki (Wielkiej Soboty, Wielkanocy i Bożego Ciała).

Z informacji zawartych w trzecim wierszu tab. 2 (kolumnie drugiej i trzeciej) wynika, że ogólny przeciętny błąd prognoz otrzymanych na podstawie modelu, w którym parametry są przeciętnymi z ocen dla podokresów poprzedzających i następnych po podokresach z lukami (LUK_SR), jest o 0,91 punktu procentowego niższy niż w modelu ze zmiennymi zero-jedynkowymi (LUK_01).

Także większość prognoz zdezagregowanych otrzymanych na podstawie modelu LUK_SR okazała się dokładniejsza w porównaniu z otrzymanymi na podstawie modelu LUK_01. Największą różnicę, wynoszącą 5,11 punktu procentowego, otrzymano dla zmiennej P8, a następnie P44 (3,08 punktu procentowego). Różnice z przedziału między 2 a 3 punktami procentowymi otrzymano dla: zmiennej P26 (2,62), poniedziałku (2,40) i czwartku (2,30). Również dokładniejsze okazały się prognozy otrzymane dla modelu LUK_SR dla zmiennych wykorzystanych do uśredniania parametrów. Najwyższe różnice otrzymano dla zmiennych: P7_9 (3,49 punktu procentowego) oraz P43_45 (2,01 punktu procentowego).

Tabela 1. Oceny parametrów strukturalnych i struktury stochastycznej równań dla pełnych danych i szeregów z lukami systematycznymi

Zmienna	Modele				Zmienna	Modele				Zmienna	Modele			
	PEL_01	PEL_SR	LUKI_01	LUK_SR		PEL_01	PEL_SR	LUKI_01	LUK_SR		PEL_01	PEL_SR	LUKI_01	LUK_SR
W. wolny	227,67	228,64	238,37	228,64	P13	–16,3		–17,8		P43	29,75		29,46	
ST	0,53	–0,71	–0,27	–1,41	P14	–5,43				P44	26,59			
LU	–6,96		–7,47		P15	4,91		3,7		P45	21,99		22,75	
MAR	–13,25				P16	12,32	12,3	11,37	10,87	P46	17,34	17,53	18,17	17,33
KW	–39,63		–42,38		P17	16,9	17,07	15,69	14,95	P47	8,71	8,61	9,07	8,68
MAJ	–66,23	–67,86	–69,59	–67,03	P18	22,02	22,37	21,72	20,74	t	1,48E–05[5]	–4,44E–05	–1,62E–05	–6,59E–05
CZE	–64,7	–65,84	–67,51	–64,91	P19	25,29		25,3		Y_{t-48}	0,4681	0,4634	0,4467	0,473
LIP	–60,03	–61,01	–62,45	–59,92	P20	26,11				N_ROK	–37,55	–57,98	–68,58	–60,64
SIE	–62,77	–64,03	–64,83	–62,51	P21	24,68		25,08		W_SOB	–49,31	–42,49	–22,68	–39,49
WRZ	–53,68	–54,68	–56,3	–54,17	P22	26,5	26,87	27,51	26,27	WIELK	–49,31	–63,69	–47,71	–59,65
PAZ	–40,58	–41,17	–41,97	–40,28	P23	28,07	28,45	28,26	26,94	S_MAJ	–37,04	–39,62	–41,13	–42,57
LIS	–13,18	–13,55	–13,78	–13,35	P24	28,37	28,61	28,8	27,47	B_C	–29,47	–33,8	–30,06	–34,33
PN	18,17		16,22		P25	27,98		28,49		WNMP	–44,99	–50,15	–47,51	–54,7
WT	1,17				P26	27,79				01_LIS	–89,07	–91,88	–89,19	–92,41
SR	1,64		1,71		P27	27,6		27,96		11_LIS	–14,49	–13,99	–13,6	–13,6
CZW	0,53	0,61	0,35	0,47	P28	26,82	27	26,89	25,64	B_NAR	–60,97	–5,33	2,19	–5,29
SOB	–24,83	–24,66	–25,34	–25,23	P29	23,73	23,71	24,25	23,18	LU_KW		–20,64		–24,56
NDZ	–42,59	–43,37	–43,4	–42,9	P30	22,42	22,52	22,7	21,67	PN_SR		7,03		9,86
P1	–7,96		–8,65		P31	19,4		19,87		P1_3		–13,06		–13,47
P2	–12,69				P32	19,26				P7_9		–27,08		–27,56
P3	–17,93		–19,46		P33	20,46		21,21		P13_15		–5,63		–6,72
P4	–21,08	–21,72	–22,68	–21,65	P34	22,48	22,73	23,57	22,47	P19_21		25,68		24,06
P5	–24,91	–25,1	–26,24	–25,05	P35	25,76	25,7	27,04	25,8	P25_27		28,18		26,93
P6	–25,7	–25,88	–26,82	–25,59	P36	26,22	26,24	27,48	26,22	P31_33		19,9		19,59
P7	–26,58		–28,52		P37	24,13		25,22		P37_39		27,51		25,67
P8	–25,77				P38	28,15				P43_45		26,46		24,87
P9	–27,38		–29,21		P39	29,29		28,65		R2	0,8565	0,8537	0,8463	0,8322
P10	–26,11	–26,62	–28,17	–26,88	P40	30,17	30,65	29,41	28,01	SE	27,53	27,8	28,61	29,88
P11	–23,49	–23,96	–24,83	–23,63	P41	30,67	30,9	29,76	28,37	Vs	7,52	7,5936	7,9404	8,2929
P12	–21,28	–21,54	–23,24	–22,15	P42	30,42	30,72	30,37	28,96	DW	0,804	0,804	0,806	0,806

Źródło: opracowanie własne.

Tabela 2. Oceny błędów prognoz interpolacyjnych

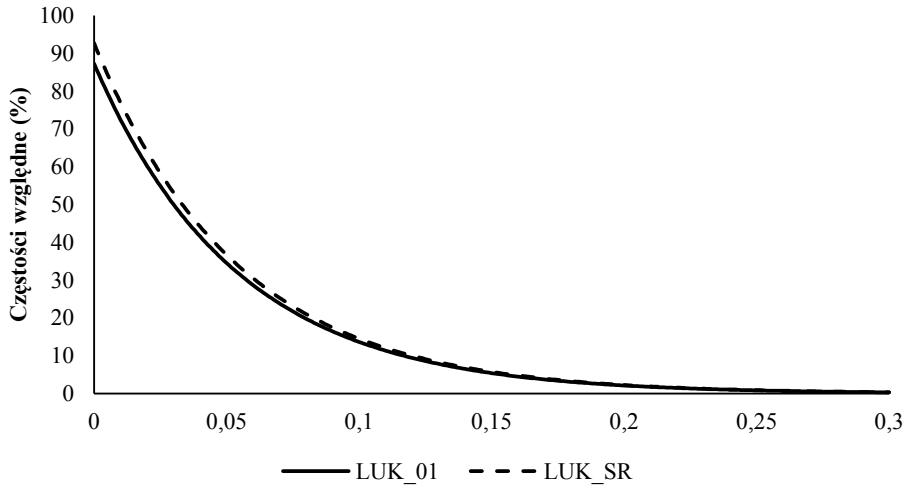
Zmienna	Modele		Zmienna	Modele		Zmienna	Modele	
	LUK_01	LUK_SR		LUK_01	LUK_SR		LUK_01	LUK_SR
OG	6,3	5,39	P10	5,29	5,75	PZ39	4,39	5,52
ST	5,98	4,69	P11	5,17	5,82	P40	4,15	5,42
LU	5,15	4,22	P12	4,74	5,71	P41	3,91	5,44
MAR	5,74	4,71	P13	4,54	5,97	P42	3,86	4,8
KW	7,04	8,18	P14	5,97	5,54	P43	3,3	4,13
MAJ	6,53	5,48	P15	5,16	4,95	P44	7,19	4,11
CZE	6,04	5,12	P16	5,14	4,89	P45	3,81	3,92
LI	6,26	5,22	P17	5,15	4,69	P46	4,51	4,47
SIE	7,66	6,47	P18	4,97	4,68	P47	4,44	4,32
WRZ	6,62	5,64	P19	4,56	4,63	P48	4,67	4,44
PAZ	6,71	5,85	P20	7,66	5,86	N_ROK	8,02	7,38
LIS	6,53	5,49	P21	4,62	4,9	W_SOB	7,47	8,89
GRUD	6,80	5,44	P22	4,59	4,76	WIELK	13,01	13,58
PN	10,10	7,7	P23	4,37	4,25	S_MAJ	8,42	7,85
WT	4,89	5,01	P24	4,41	4,43	BC	10,67	12,66
ŚR	7,07	5,11	P25	4,33	4,43	WNMP	6,49	7,1
CZW	6,95	4,65	P26	7,65	5,03	LIS_1	7,02	9,83
PT	6,67	4,81	P27	4,2	4,38	LIS_11	4,83	4,95
SOB	6,03	5,52	P28	4,61	4,8	B_NAR	8,15	7,28
NDZ	7,02	6,94	P29	4,68	4,79	LU_KW	5,86	5,23
P1	5,22	4,97	P30	4,84	4,88	PN_SR	6,07	5,45
P2	8,4	7,1	P31	4,69	4,37	P1_3	7,51	6,62
P3	5,68	6,04	P32	6,12	5,1	P7_9	10,19	6,7
P4	5,66	5,69	P33	5,32	4,65	P13_15	5,63	5,52
P5	5,55	5,52	P34	5,59	4,79	P19_21	6,73	5,53
P6	5,55	5,65	P35	5,99	5,36	P25_27	6,63	4,84
P7	5,68	5,76	P36	6,49	5,83	P31_33	5,78	4,92
P8	12,34	7,23	P37	5,52	5,42	P37_39	6,85	6,19
P9	4,79	5,21	P38	7,67	6,5	P43_45	6,09	4,08

Źródło: opracowanie własne.

Spośród świąt najmniej dokładne prognozy otrzymano dla Wielkiejnocy i Bożego Ciała. Dokładniejsze prognozy na podstawie modelu ze zmiennymi zero-jedyn-

kowymi otrzymano dla większości świąt. Największą różnicę dodatnią otrzymano dla Wszystkich Świętych (2,80 punktu procentowego), Święta Bożego Ciała (1,99 punktu procentowego). Spośród 48 okresów półgodzinnych jedynie dla 15 prognozy otrzymane na podstawie modelu LUK_01 były dokładniejsze, przy czym tylko dla 4 zmiennych P13 oraz P39-P41 różnice ocen błędów przekraczały 1 punkt procentowy. Największą różnicą charakteryzowała się zmienna P41(1,53 punktu procentowego).

Na rysunku 2 przedstawiono w postaci graficznej kształtowanie się aproksymant rozkładów empirycznych błędów prognoz interpolacyjnych. Rozkłady empiryczne najlepiej opisywał rozkład Gamma.



Rys. 2. Rozkłady błędów prognoz interpolacyjnych

Źródło: opracowanie własne.

Z rysunku 2 wynika, że rozkłady błędów są silnie prawostronnie asymetryczne, przy czym bardziej dokładne są prognozy otrzymane na podstawie modelu LUK_SR.

W tabeli 3 zestawione zostały oceny błędów prognoz ekstrapolacyjnych otrzymanych na podstawie modeli dla pełnych danych i luk systematycznych.

Z porównania ocen błędów prognoz ogółem wynika, że błędy zarówno dla pełnych danych, jak i szeregów z lukami przyjęły niższe wartości dla prognoz otrzymanych na podstawie modeli PEL_SR oraz LUK_SR, tzn. modeli, w których dokonano uśredniania parametrów. Różnice w dokładności prognoz są niższe niż dla prognoz interpolacyjnych i kształtują się na poziomie odpowiednio: 0,10 i 0,27 punktu procentowego. Relacja dokładności prognoz jest jednak odmienna od relacji własności predyktywnych, kiedy to nieco lepsze były modele ze zmiennymi zero-jedynkowy-

Tabela 3. Oceny błędów prognoz ekstrapolacyjnych

Zmienna	Modele				Zmienne	Modele				Zmienne	Modele			
	PEL_01*)	PEL_SR	LUK_01	LUK_SR		PEL_01*)	PEL_SR	LUK_01	LUK_SR		PEL_01*)	PEL_SR	LUK_01	LUK_SR
OG	5,26	5,16	5,74	5,47	P10	7,18	6,72	7,32	7,30	P39	5,84	5,83	5,75	5,97
ST	4,24	4,29	4,64	4,49	P11	6,94	6,45	7,18	7,12	P40	5,68	5,83	5,59	5,94
LU	5,04	3,55	5,25	3,65	P12	6,46	5,96	6,57	6,45	P41	5,12	5,27	5,08	5,39
MAR	5,56	4,73	8,20	4,72	P13	6,07	7,81	6,25	8,31	P42	4,88	4,69	4,97	4,90
KW	5,36	8,95	5,30	8,67	P14	5,51	4,89	7,23	5,24	P43	4,50	4,03	4,58	4,18
MAJ	6,05	5,80	6,36	6,27	P15	5,60	3,83	5,64	3,93	P44	4,25	3,94	6,02	4,20
CZE	6,23	5,61	6,30	6,24	P16	5,76	4,96	5,82	5,19	P45	4,29	4,37	4,47	4,70
LI	5,54	5,02	5,75	5,71	P17	5,48	4,64	5,42	4,78	P46	4,37	4,08	4,60	4,42
SIE	4,75	4,48	5,31	5,01	P18	4,90	4,20	4,97	4,46	P47	4,94	4,60	5,18	5,00
WRZ	5,48	4,88	5,64	5,33	P19	4,36	3,85	4,46	4,11	P48	4,54	4,84	4,91	4,90
PAZ	4,16	4,32	4,73	4,64	P20	3,97	3,42	5,08	3,65	N_ROK	10,55	9,16	8,91	9,16
LIS	4,73	4,30	4,99	4,60	P21	3,97	3,70	4,13	3,97	W_SOB	13,03	29,45	8,05	4,00
GRUD	4,00	6,06	6,28	6,20	P22	3,92	3,65	4,15	4,02	W_NOC	10,83	18,29	8,27	17,44
PN	9,88	5,04	9,61	5,21	P23	3,96	4,03	3,62	3,85	MAJ	6,94	7,42	8,07	7,32
WT	4,99	5,67	5,26	6,46	P24	3,80	3,53	3,93	3,82	BC	2,57	4,44	3,60	4,11
ŚR	4,76	5,44	5,40	6,31	P25	3,84	3,59	3,99	3,86	WNMP	5,76	6,39	6,26	6,23
CZW	4,71	5,61	5,36	6,41	P26	3,69	3,52	5,65	3,78	LIS_1	9,77	11,95	9,79	11,76
PT	4,63	3,98	4,96	4,12	P27	3,68	3,53	3,80	3,77	LIS_11	11,22	5,25	10,78	5,81
SOB	3,52	4,17	4,36	4,00	P28	3,40	3,33	3,50	3,51	B_NAR	9,38	3,94	6,87	4,43
NDZ	5,16	6,29	5,62	5,95	P29	3,68	6,97	3,86	7,58	LU_KW	5,36	5,78	6,30	5,71
P1	6,31	5,59	6,51	5,96	P30	3,40	6,34	3,57	6,86	PN_SR	4,79	5,57	5,34	6,39
P2	6,78	6,43	9,33	6,93	P31	3,62	3,47	3,82	3,76	P1_3	6,79	6,43	7,65	6,91
P3	6,92	7,28	7,10	7,83	P32	3,95	5,05	4,46	4,44	P7_9	7,39	6,88	9,78	7,41
P4	7,08	6,60	7,31	7,19	P33	4,44	5,38	4,71	4,74	P13_15	5,73	5,51	6,37	5,83
P5	7,07	6,78	7,33	7,37	P34	4,97	4,47	5,30	4,77	P19_21	4,09	3,66	4,56	3,91
P6	7,23	6,85	7,53	7,50	P35	5,74	5,20	6,11	5,55	P25_27	3,73	3,54	4,48	3,80
P7	7,16	6,78	7,34	7,31	P36	5,54	5,35	5,87	5,68	P31_33	3,57	5,59	3,75	6,06
P8	7,51	7,01	14,72	7,54	P37	5,65	5,88	5,92	6,04	P37_39	5,74	5,88	5,81	6,02
P9	7,10	6,86	7,28	7,39	P38	5,91	5,92	5,76	6,06	P43_45	4,37	4,11	5,02	4,36

Źródło: opracowanie własne; *) [Szmuksta-Zawadzka, Zawadzki 2011].

mi. Oznacza to, że nie zawsze równania o lepszych charakterystykach ocen parametrów struktury stochastycznej dają prognozy o niższych błędach.

W następnych wierszach i kolumnach zestawione zostały oceny błędów prognoz zdezagregowanych na miesiące, dni tygodnia oraz okresy półgodzinne i święta. W ostatnich 10 wierszach zamieszczono oceny błędów dla podokresów, dla których zostały wyznaczone średnie wartości parametrów.

Z analizy błędów prognoz zdezagregowanych dla danych z lukami wynika, że w wielu przypadkach różnice w dokładności są znacznie większe w porównaniu z różnicami błędów ogółem. W większości przypadków dokładniejsze okazały się prognozy otrzymane na podstawie modelu z uśrednionymi parametrami (LUK_SR). Dla rocznego cyklu wahań dla 11 miesięcy dokładniejsze prognozy otrzymano na podstawie modelu LUK_SR. Największe różnice otrzymano dla marca i lutego – wyniosły one odpowiednio: 3,48 oraz 1,60 punktu procentowego. Jedynie dla kwietnia oceną niższą o 3,36 punktu charakteryzowały się prognozy otrzymane na podstawie modelu LUK_01. Odmiennie sytuacja kształtuje się dla cyklu tygodniowego – dla czterech dni niższe oceny otrzymano na podstawie modelu ze zmiennymi zero-jedynkowymi. Różnice ocen błędów dla trzech dni (wtorku, środy, czwartku) kształtowały się na poziomie 1 punktu procentowego. Jednak największą różnicę wynoszącą 4,40 punktu otrzymano dla poniedziałku, przy czym oceną niższą charakteryzowały się prognozy otrzymane na podstawie modelu LUK_SR. Dla cyklu dobowego spośród 48 okresów półgodzinnych dla 26 dokładniejsze prognozy otrzymano na podstawie modelu LUK_01. Jednak tylko dla zmiennej P13 różnica ocen wynosząca 2,06 była wyższa od 1 punktu procentowego. W przypadku prognoz otrzymanych na podstawie modelu LUK_SR miało to miejsce w 5 przypadkach. Największe różnice ocen błędów prognoz między modelami LUK_SR oraz LUK_01 otrzymano dla zmiennych P8, P2 oraz P14 – wyniosły one odpowiednio: 7,18; 2,40 i 1,99 punktu procentowego. Dla dwóch dni świątecznych: Wielkiej Soboty oraz Wielkanocy, otrzymano na podstawie niektórych modeli prognozy niedopuszczalne, ponieważ błędy prognoz znacznie przekraczały przyjęte kryterium dopuszczalności prognoz wynoszące 10%. W przypadku święta Wszystkich Świętych dla modeli ze zmiennymi zero-jedynkowymi przekroczenie to było znacznie niższe i wynosiło mniej niż 2 punkty procentowe. Oceny błędów prognoz otrzymanych na podstawie obu modeli dla pozostałych dni świątecznych były bardzo zbliżone – różnice nie przekraczały 0,85 punktu procentowego.

Natomiast dość znacznie różniły się one poziomem. Najniższe oceny błędów otrzymano dla Bożego Ciała i Bożego Narodzenia – kształtowały się one w przedziale od 3,94 do 4,44%. Różnice ocen błędów niższe od 1 punktu otrzymano dla zmiennych grupujących, w skład których wchodziły podokresy użyte do uśredniania parametrów. W 9 przypadkach na 10 nieco dokładniejsze prognozy otrzymano na podstawie modelu LUK_01.

4. Podsumowanie

Z przeprowadzonych w pracy rozważań oraz przykładu empirycznego będącego ich ilustracją można wyprowadzić następujące wnioski:

1. Modele ze zmiennymi zero-jedynkowymi (LUK_01) oraz uśrednionymi parametrami szacowane na podstawie szeregów z lukami systematycznymi (LUK_SR) obejmującymi ok. 35% ich długości charakteryzowały się bardzo zbliżonymi własnościami predyktywnymi do odpowiadających ich modelom dla danych pełnych (PEL_01 i PEL_SR).

2. Nieznacznie lepszymi własnościami predyktywnymi charakteryzowały się modele ze zmiennymi zero-jedynkowymi.

3. Relacja ocen średnich względnych błędów prognoz ogółem, zarówno inter-, jak i ekstrapolacyjnych była odwrotna, tzn. niższe ich oceny otrzymano dla modeli z uśrednionymi parametrami.

4. Oceny błędów prognoz dla okresów, w których wystąpiły luki systematyczne, były tylko nieznacznie wyższe w porównaniu z ocenami dla modeli szacowanych na podstawie pełnych danych.

5. Spośród dwóch modeli dla danych z lukami nieco wyższą dokładnością charakteryzował się model LUK_SR, przy czym różnica ocen błędów była wyższa dla prognoz interpolacyjnych.

6. Dokonując rekapitulacji wyników osiągniętych w pracy, możemy powiedzieć, że modele szeregu czasowego mogą być użyteczne w prognozowaniu brakujących danych w szeregach o bardzo wysokiej częstotliwości obserwowania, także w przypadku występowania luk systematycznych.

Literatura

Dittmann P., *Prognozowanie w przedsiębiorstwie. Metody i ich zastosowanie*, Wolters Kluwer Polska, Kraków 2008.

Szmuksta-Zawadzka M., Zawadzki J., *Modelowanie predyktywne i prognozowanie na podstawie „oszczędnych" modeli szeregu czasowego z wahaniami sezonowymi*, „Przegląd Statystyczny" 1998, z. 3, s. 381-386.

Szmuksta-Zawadzka M., Zawadzki J., *Prognozowanie brakujących informacji a modele oszczędne dla okresowych szeregów czasowych,* [w:] *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, Materiały 21 Ogólnopolskiego Seminarium Naukowego, Kraków 2000, s. 123-129.

Szmuksta-Zawadzka M., Zawadzki J., *Forecasting on the basis of "Parsimonius" hierarchical models*, Dynamic Econometric Models, vol. 7, Toruń Nikolaus Copernicus University, 2006, s. 37-47.

Szmuksta-Zawadzka M., Zawadzki J., *Zastosowanie modelowania ekonometrycznego w prognozowaniu brakujących danych w szeregach o wysokiej częstotliwości*, Ekonometria [Econometrics] nr 34, Wrocław 2011, s. 303-313.

Zeliaś A., Pawełek B., Wanat S., *Prognozowanie ekonomiczne. Teoria, przykłady, zadania,* Wydawnictwo Naukowe PWN, Warszawa 2003.

ABOUT A METHOD OF FORECASTING OF MISSING DATA IN THE HIGH FREQUENCY TIME SERIES WITH SYSTEMATIC GAPS

Summary: This paper attempts to extend the considerations contained in the article [Szmuksta-Zawadzka, Zawadzki, 2011] concerning forecasting for high frequency time series with unsystematic gaps in the case of the occurrence of systematic gaps. Theoretical considerations will be illustrated with an empirical example of modelling and forecasting the demand of electrical energy in half-an-hour periods in one of urban areas.

Keywords: high-frequency time series, forecasting, missing data, systematic gaps.