

SYMULACYJNE BADANIE WPLYWU WYSTĘPOWANIA LUK SYSTEMATYCZNYCH W SZEREGU CZASOWYM DLA DANYCH DZIENNYCH NA DOKŁADNOŚĆ PROGNOZ

Maciej Oesterreich

Wydział Ekonomiczny

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

e-mail: maciej.oesterreich@zut.edu.pl

Streszczenie: W pracy przedstawiono wyniki symulacyjnej analizy wpływu występowania luk systematycznych na dokładność prognoz inter- i ekstrapolacyjnych w dziennych szeregach czasowych. Do budowy prognoz wykorzystano klasyczny model szeregu czasowego, w którym wahania sezonowe o cyklach: tygodniowym i rocznym, były opisane za pomocą zmiennych zero-jedynkowych. Zmienną, którą poddano analizie, była dzienna sprzedaż paliw płynnych na stacji paliw X w latach 2012-2014. Pierwsze trzydzieści miesięcy stanowiło przedział czasowy próby, a ostatnie sześć było okresem empirycznej weryfikacji prognoz. Rozpatrywanych było jedenaście wariantów luk systematycznych. Obliczenia zostały wykonane z wykorzystaniem pakietu R oraz Statistica 12.

Słowa kluczowe: zmienne o wysokiej częstotliwości obserwowania, złożone wahania sezonowe, luki systematyczne, prognozowanie, modele szeregu czasowego.

WSTĘP

W pracy [Oesterreich 2017] przedstawiono wyniki symulacyjnej analizy wpływu występowania luk niesystematycznych na dokładność prognoz inter- i ekstrapolacyjnych w szeregu czasowym dla danych dziennych. Wskazano w nim na podobieństwa, w kształcie i kierunku zmian rozkładów średnich względnych błędów (MAPE) obu rodzajów prognoz, do otrzymanych na podstawie danych miesięcznych.

Nasuwa się zatem pytanie, czy podobne wnioski będzie można wyciągnąć również w odniesieniu do luk systematycznych? Czy rozkłady błędów prognoz

inter- i ekstrapolacyjnych oszacowanych dla danych dziennych, co do kierunku i kształtu, będą zbliżone do otrzymanych dla danych miesięcznych? Celem niniejszej pracy jest więc rozszerzenie badań na przypadek wystąpienia luk systematycznych.

W obliczeniach wykorzystano metody symulacyjne, za pomocą których wygenerowano jedenaście wariantów, różniących się z góry zadaną liczbą luk odnoszących się do wahań o cyklu: 7-dniowym lub/i 12-miesięcznym. Każdy z wariantów zawierał wszystkie możliwe kombinacje układów luk. Dane te posłużyły następnie do oszacowania modeli, wyznaczenia prognoz inter- i ekstrapolacyjnych oraz obliczania ich średnich względnych błędów (MAPE) (zob. [Oesterreich 2012]).

METODY BADAWCZE

Do budowy prognoz inter- i ekstrapolacyjnych wykorzystany został klasyczny addytywny model szeregu czasowego, w którym zarówno tygodniowa jak i roczna sezonowość była opisana za pomocą zmiennych zero-jedynkowych. Zapis analityczny modelu w postaci addytywnej jest następujący:

$$Y_t = \alpha_0 + \alpha_1 t + \sum_{i=1}^7 D_i Q_{idt} + \sum_{j=1}^{12} M_j Q_{jmt} + \varepsilon_t, \quad t = 1, 2, \dots, n \quad (1)$$

gdzie:

Y_t – zmienna objaśniana,

α_0, α_1 – parametry trendu liniowego,

t – zmienna czasowa,

Q_{idt} – zmienna przyjmująca 1 dla i -tego dnia tygodnia oraz 0 dla pozostałych,

Q_{jmt} – zmienna przyjmująca 1 dla j -tego miesiąca roku oraz 0 dla pozostałych,

D_i, M_j – parametry strukturalne,

ε_t – składniki losowe.

Do wyboru statystycznie istotnych składowych wykorzystano procedurę regresji krokowej opartej o statystykę F [Anderson i in. 2011, s. 730]. Minimalny poziom wejścia zmiennej do modelu ustalono na $F=1$. W obliczeniach wykorzystano autorski algorytm *reg_tree_alg* w środowisku R.

W pracy przeanalizowano jedenaście wariantów luk systematycznych. Liczba ta wynika z faktu występowania w analizowanej zmiennej dwóch typów wahań – tygodniowych i rocznych. Oznacza to, że liczba wariantów uległa podwojeniu w porównaniu do analiz prowadzonych na podstawie szeregów miesięcznych [Oesterreich 2014b].

Badaniu poddano warianty, w których z pełnego szeregu usunięto jeden, dwa lub trzy dni tygodnia, a w przypadku miesięcy jeden lub dwa miesiące w roku. W dalszej części artykułu warianty luk zostaną oznaczone w sposób skrótowy np.

D1_M2 oznacza wariant, w którym usunięto jeden dzień w tygodniu oraz dwa miesiące w roku. W tabeli 1 zestawione zostały liczby kombinacji układów luk dla poszczególnych wariantów.

Tabela 1. Liczby kombinacji luk dla 11 rozpatrywanych wariantów luk systematycznych

		Miesiące		
		M0	M1	M2
Dni	D0	-	12	66
	D1	7	84	462
	D2	21	252	1386
	D3	35	420	2310

Źródło: opracowanie własne

Symbole D0 i M0 oznaczają, że luki występowały odpowiednio tylko: w miesiącach lub dniach tygodnia.

Wyznaczone, w toku procedury symulacyjnej dla każdego z wariantów, kombinacje układów luk różniły się ich odsetkami w poszczególnych dniach tygodnia i miesiącach. Informacje o minimalnych i maksymalnych wielkościach odsetkach luk w poszczególnych wariantach przedstawiono w tabeli 2.

Tabela 2. Minimalne i maksymalne odsetki luk w poszczególnych wariantach luk (w %)

		Miesiące					
		M0		M1		M2	
		min	max	min	max	min	max
Dni	D0	-	-	6,57	10,20	13,16	20,39
	D1	14,25	14,36	19,74	23,25	25,44	32,02
	D2	28,51	28,72	33,00	36,29	37,61	43,64
	D3	42,76	42,98	46,27	49,12	49,89	55,15

Źródło: opracowanie własne

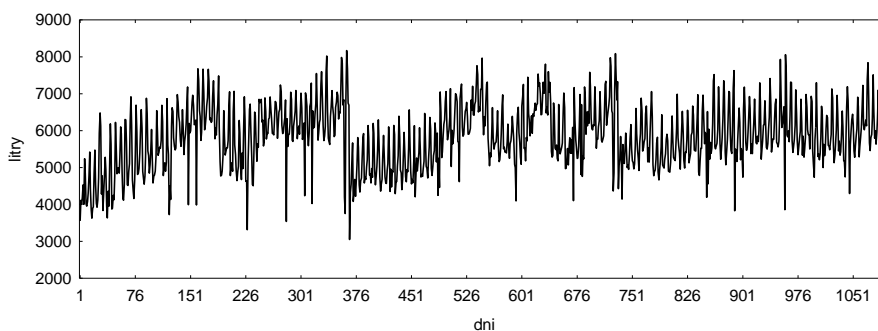
Z informacji zawartych w tabeli 2 wynika, że minimalną wielkość odsetka luk, wynoszącą 6,57%, otrzymano dla wariantu D0_M1 a maksymalną, wynoszącą 55,15%, dla wariant D3_M2. Najwyższymi różnicami odsetek luk charakteryzowały się warianty, w których usunięto dwa miesiące: D0_M2, D1_M2 i D2_M2 i D3_M2. Wynosiły one odpowiednio: 7,23, 6,58, 6,03 i 5,26 punktu procentowego (p.p.).

PRZYKŁAD EMPIRYCZNY

Modelowaniu i prognozowaniu poddano zmienną dotyczącą dziennej sprzedaży paliw płynnych na stacji paliw X w latach 2012-2014 (w litrach). Obejmowała ona łączną sprzedaż: benzyny bezołowiowej Pb95 oraz Pb98, a także oleju napędowego. Dane statystyczne za okres od 1 stycznia 2012 do 30 czerwca 2014 (912 obserwacji) stanowiły przedział czasowy „próby”. Okresem

empirycznej weryfikacji prognoz było drugie półrocze 2014 roku (184 obserwacje). Na rysunku 1 przedstawiano kształtowanie się zmiennej prognozowanej.

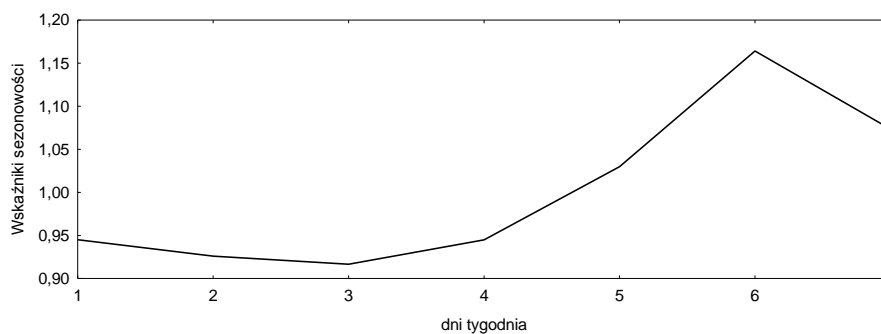
Rysunek 1. Sprzedaż paliw płynnych na stacji paliw X w okresie od 1.01.2012-31.12.2014



Źródło: opracowanie własne

Na podstawie szeregu czasowego oszacowano wskaźniki sezonowości dla poszczególnych dni tygodnia oraz miesięcy. Ich kształtowanie przedstawiono w postaci graficznej na rysunkach 2 i 3.

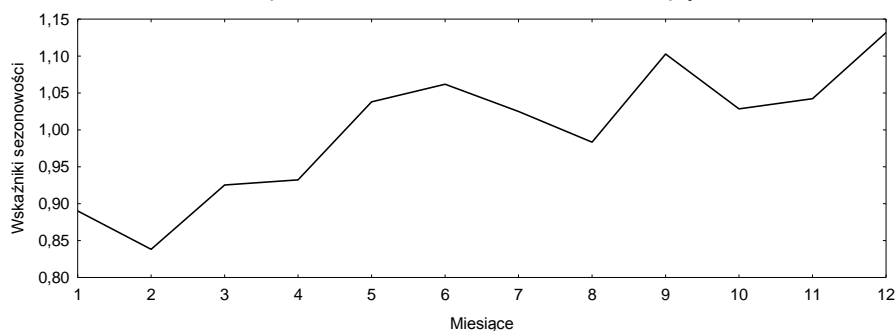
Rysunek 2. Kształtowanie się wskaźników sezonowości dla dni tygodnia



Źródło: opracowanie własne

Z rysunku 2 wynika, że najwyższą ocenę wskaźniki sezonowości dla dni tygodnia, przyjęły w sobotę (ok. 116%), natomiast najniższy jego poziom odnotowano dla środy (ok. 92%). Różnica pomiędzy skrajnymi ich wartościami wyniosła ok. 24 p.p.

Rysunek 3. Kształtowanie się wskaźników sezonowości dla miesięcy



Źródło: opracowanie własne

Z rysunku 3 wynika, że w przebiegu ocen wskaźników sezonowości możemy wyróżnić trzy lokalne maksima sezonowe w: czerwcu (ok. 106%), wrześniu (ok. 110%) i grudniu (ok. 113%). Minimalną ocenę wskaźnik sezonowości przyjął w lutym (ok. 84%). Różnica pomiędzy ocenami wskaźników w miesiącach o skrajnych wartościach wynosiła ok. 29 p.p. i była ok. 5 p.p. wyższe niż dla dni tygodnia.

W procesie wyboru składowych dla modeli z wahaniami sezonowymi, których parametry szacowane są na podstawie szeregów zawierających luki systematyczne można natknąć się na następujące problemy [Zawadzki 2003, s.26], [Oesterreich 2014a]:

1. współliniowość zmiennych objaśniających (składowych),
2. zerowe wariancje niektórych z nich,
3. liniowe kombinacje części składowych.

Wystąpienie kombinacji liniowych powoduje konieczność budowy więcej niż jednego modelu dla danego układu luk. Najczęściej z tego rodzaju sytuacją mamy do czynienia w przypadku klasycznych modeli szeregu czasowego, w których sezonowość opisywana jest za pomocą wielomianu trygonometrycznego [Zawadzki 1999], [Oesterreich 2014a], ale występuje ona również w przypadku modeli, których sezonowość opisano za pomocą zmiennych zero-jedynkowych [Oesterreich 2014b].

W tabeli 3 przedstawiano informację na temat liczby modeli oszacowanych

Tabela 3. Liczba oszacowanych modeli w zależności od wariantu luk

		Miesiące		
		M0	M1	M2
Dni	D0	-	12 (12)	66 (66)
	D1	7 (7)	84 (84)	466 (462)
	D2	23 (21)	267 (252)	1449 (1386)
	D3	44 (35)	506 (420)	2739 (2310)

* w nawiasach podano liczbę kombinacji luk dla poszczególnych wariantów

Źródło: opracowanie własne

dla każdego z analizowanych wariantów luk. Liczby w nawiasach odpowiadają liczbom kombinacji układów luk.

Z informacji zawartych w tabeli 3 wynika, że w siedmiu z jedenastu wariantów wystąpiły kombinacje liniowe składowych, dla których oszacowano więcej niż jeden model. Najmniejszą ich liczbę odnotowano w wariacie D2_M0 – 2, a największą w wariacie D3_M2 – 429. Największym procentowym udziałem takich kombinacji charakteryzował się wariant D3_M0 – 25,7%.

W tabeli 4 przedstawiono informację o wybranych statystykach opisowych rozkładów średnich względnych błędów (MAPE) prognoz interpolacyjnych.

Tabela 4. Wybrane statystyki opisowe rozkładów średnich względnych błędów prognoz interpolacyjnych

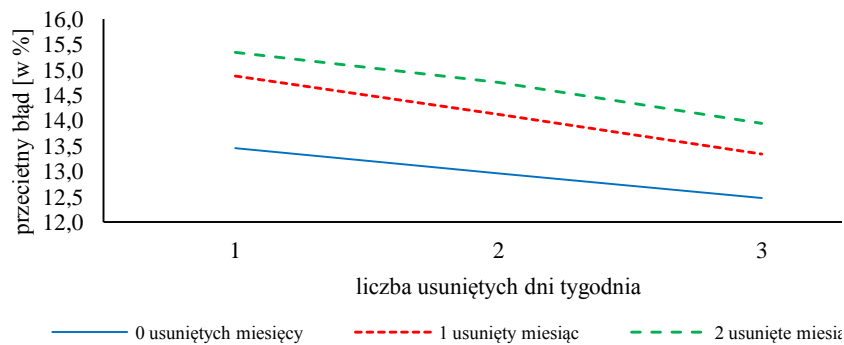
Wariant luk	\bar{x}	Me	Q1	Q3	Min	Max	Skośność	S	V _s
D0_M1	15,712	12,242	10,335	20,181	5,821	33,270	1,084	8,196	52,164
D0_M2	15,775	14,548	10,682	20,177	7,588	31,010	0,611	5,787	36,685
D1_M0	13,459	14,160	9,334	16,603	8,276	17,470	-0,641	3,450	25,632
D1_M1	14,875	14,346	12,232	16,973	7,538	26,360	0,588	4,131	27,769
D1_M2	15,341	14,652	12,313	17,943	7,687	28,144	0,553	3,991	26,015
D2_M0	12,962	12,375	11,309	15,397	8,892	17,056	0,432	2,112	16,291
D2_M1	14,121	13,788	11,787	15,748	8,385	24,069	0,719	2,858	20,237
D2_M2	14,753	14,342	12,295	16,666	8,271	26,614	0,697	3,099	21,004
D3_M0	12,471	12,492	10,666	13,870	9,817	16,253	0,205	1,897	15,213
D3_M1	13,343	13,265	11,385	14,745	9,493	22,258	0,656	2,366	17,734
D3_M2	13,943	13,661	12,031	15,422	9,222	24,783	0,768	2,627	18,843

Źródło: opracowanie własne

Z informacji zawartej w tabeli 4 wynika, że najniższą ocenę przeciętnego błędu względnego prognoz interpolacyjnych, wynoszącą 12,471%, odnotowano dla wariantu D3_M0, a najwyższą dla wariantu D0_M2 – 15,775%. Tylko nieznacznie niższy błąd – o 0,063 p.p. – odnotowano dla wariantu D0_M1. Oznacza to, że dla wariantów o najniższych odsetkach luk otrzymano najwyższe oceny przeciętnych błędów.

Taka sytuacja może wynikać z dwóch głównych przyczyn. Po pierwsze wielkość przeciętnych ocen błędów prognoz interpolacyjnych zależała przede wszystkim od rozmieszczenia luk. Z analizy zarówno ocen przeciętnych wartości rozkładów błędów jak i ich median wynika, że to liczba usuniętych dni tygodnia i miesięcy determinowała wielkość błędów. Na rysunku 4 przedstawiano kształtowanie się wartości przeciętnych błędów w zależności od liczby usuniętych dni tygodnia i miesięcy.

Rysunek 4. Wartości przeciętnych błędów względnych prognoz interpolacyjnych w zależności od liczby usuniętych dni tygodnia oraz miesięcy

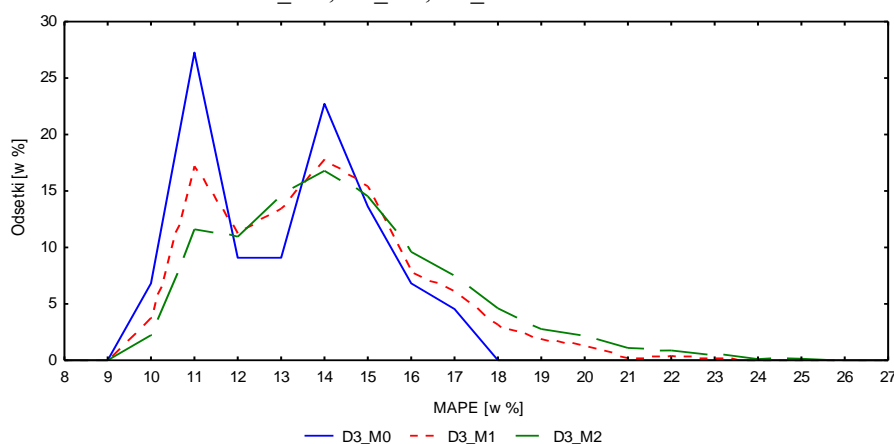


Źródło: opracowanie własne

Z rysunku wynika, że przeciętne błędy względne prognoz interpolacyjnych wykazują tę samą tendencję malejącą wraz ze wzrostem liczby luk występujących w dniach tygodnia. Różnią się one natomiast co do poziomu – rosną one wraz ze wzrostem liczby luk w miesiącach, przy czym przyrosty błędów nie przekraczają 1,5 p.p. Powyższy wynik nie jest niespodziewany, ponieważ, jak wynika z tabeli 3, równocześnie rośnie liczba kombinacji układów luk. Występowanie względnie wyższych ocen błędów dla małej liczby kombinacji związane jest z występowaniem luk, w których występują minima lub maksima sezonowe. W miarę wzrostu liczby układów luk ich wpływ maleje. Ponadto na dokładność prognoz mogły wpływać obserwacje odstające i wpływowe, których występowanie w analizowanym szeregu zostało potwierdzone [Oesterreich 2017].

Na rysunku 5 przedstawiano kształtowanie empirycznych rozkładów błędów

Rysunek 5. Empiryczne rozkłady średnich względnych błędów prognoz interpolacyjnych dla wariantów D3_M0, D3_M1, D3_M2



Źródło: opracowanie własne

prognoz interpolacyjnych dla trzech wariantów: D3_M0, D3_M1 oraz D3_M2. Warianty te charakteryzowały się największymi odsetkami luk.

Z rysunku 5 wynika, że analizowane rozkłady błędów były wielomodalne. Wraz ze wzrostem liczby luk wzrastała ocena rozstępu pomiędzy maksymalnymi a minimalnymi ocenami błędów. Jednocześnie zmniejszeniu uległ odsetek obserwacji w przedziałach na które przypadają „szczyty”. Z analizy statystyk opisowych analizowanych rozkładów (tabela 4) wynika, że rozkłady powoli przesunęły się w prawą stronę osi odciętych. Wskazują na to rosnące wartości median analizowanych rozkładów. Wykazujące wzrostową tendencję oceny współczynników zmienności sugerują, że odchylenia standardowe błędów prognoz przyrastały szybciej niż przeciętny błąd. Wynika to z faktu pojawiania się coraz większej liczby kombinacji w poszczególnych wariantach luk, dla których otrzymano oceny błędów wyraźnie wyższych od średniej.

W tabeli 5 zestawione zostały wybrane statystyki opisowe rozkładów średnich względnych błędów (MAPE) prognoz ekstrapolacyjnych.

Tabela 5. Wybrane statystyki opisowe rozkładów średnich względnych błędów prognoz ekstrapolacyjnych

Wariant luk	\bar{x}	Me	Q1	Q3	Min	Max	Skośność	S	V _s
D0_M1	8,662	8,440	8,114	9,237	7,917	10,081	0,922	0,698	8,055
D0_M2	8,861	8,705	8,037	9,580	6,570	11,347	0,457	1,002	11,311
D1_M0	9,329	9,424	9,009	9,606	8,857	9,757	-0,208	0,352	3,778
D1_M1	9,613	9,529	9,065	10,018	8,321	11,804	0,687	0,775	8,063
D1_M2	9,812	9,670	9,008	10,572	7,005	13,214	0,500	1,077	10,978
D2_M0	9,979	10,152	9,104	10,503	8,870	10,954	-0,407	0,705	7,067
D2_M1	10,272	10,207	9,603	10,926	8,578	13,260	0,403	0,984	9,581
D2_M2	10,474	10,300	9,563	11,251	7,672	14,793	0,547	1,244	11,879
D3_M0	10,368	10,904	9,017	11,363	8,545	11,861	-0,411	1,189	11,466
D3_M1	10,610	10,818	9,340	11,556	7,981	14,391	0,079	1,375	12,961
D3_M2	10,788	10,762	9,523	11,787	7,553	16,025	0,358	1,566	14,520

Źródło: opracowanie własne

Z analizy zawartych w tabeli 5 statystyk opisowych rozkładów błędów prognoz ekstrapolacyjnych wynika, że najniższą oceną przeciętną, wynoszącą 8,662%, charakteryzował się wariant D0_M1 a najwyższą – 10,788% wariant D3_M2. Oceny te były więc wyraźnie niższe od przeciętnych ocen błędów prognoz interpolacyjnych.

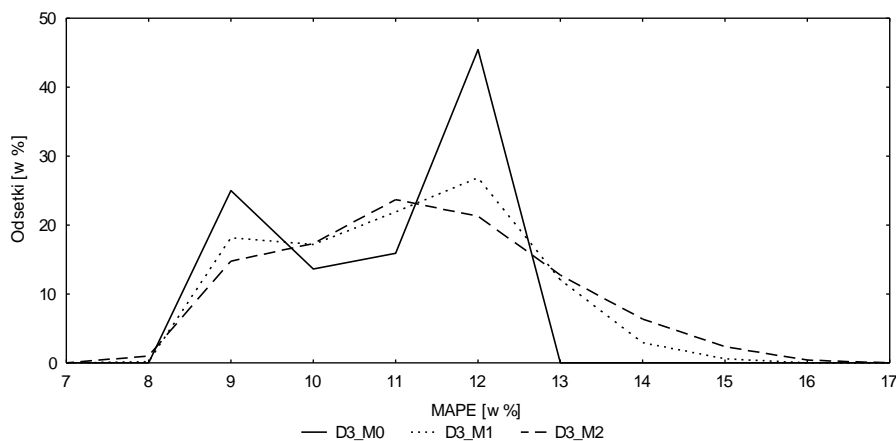
Z porównania informacji zawartych w tabeli 2 i tabeli 4 wynika, że odmiennie niż w przypadku ocen błędów prognoz interpolacyjnych, na oceny błędów prognoz ekstrapolacyjnych wpływały przede wszystkim odsetki luk w wariantach. Ich wzrost negatywnie wpływał na własności predyktywne wykorzystanego modelu. Należy jednak wskazać na niski, wynoszący tylko 2,126 p.p., rozstęp pomiędzy przeciętnymi błędami dla wariantów o skrajnych udziałach luk. Może to być spowodowane, podobnie jak w przypadku błędów prognoz interpolacyjnych, tym,

że część z wyeliminowanych obserwacji stanowiły obserwacje wpływowe i odstające. Ich usunięcie poprawiło własności predyktorów.

Z analizy median rozkładów błędów prognoz ekstrapolacyjnych dla wariantów wynika, że kierunek ich zmian był podobny do zmian wartości przeciętnych błędów. Jedynie dla wariantów z których usunięto trzy dni tygodnia, kierunek ten był odwrotny.

Na rysunku 2 przedstawiano kształtowanie się empirycznych rozkładów błędów prognoz interpolacyjnych dla trzech wariantów: D3_M0, D3_M1 oraz D3_M2.

Rysunek 6. Empiryczne rozkłady średnich względnych błędów prognoz ekstrapolacyjnych dla wariantów D3_M0, D3_M1, D3_M2



Źródło: opracowanie własne

Z analizy rysunku 6 oraz informacji zamieszczonych w tabeli 5 wynika, że empiryczne rozkłady błędów względnych prognoz ekstrapolacyjnych, podobnie jak prognoz interpolacyjnych, charakteryzowały się wielomodalnością. Tylko w przypadku rozkładu dla wariantu D3_M2 możemy mówić o jego jednomodalnym charakterze. Wyraźnemu zwiększeniu ulega rozstęp między maksymalnymi a minimalnymi ocenami błędów w poszczególnych wariantach. Z analizy ocen odchyłeń standardowych wynika, że podobnie jak to miało miejsce w przypadku rozkładów błędów prognoz interpolacyjnych, przyrastały one szybciej niż błąd przeciętny. Wpływało to bezpośrednio na wielkość ocen współczynników zmienności.

PODSUMOWANIE

Z analiz przeprowadzonych w pracy można wyciągnąć następujące wnioski:

1. Empiryczne rozkłady średnich względnych błędów prognoz interpolacyjnych i ekstrapolacyjnych charakteryzowały wielomodalnością i nieregularnym

- przebiegiem. Zbliżony kształt empirycznych rozkładów błędów prognoz obserwowano dla danych miesięcznych: dla modelu klasycznego, którego stała sezonowość opisana była za pomocą wielomianu trygonometrycznego [Oesterreich 2014a, rysunek 3] oraz modelu hierarchicznego H62 [Oesterreich 2014b, tabela 4]. Różniły się one jednak wyraźnie od rozkładów błędów prognoz dla luk niesystematycznych (zob. [Oesterreich 2012, 2017]).
2. Kierunek zmian statystyk opisowych rozkładów średnich względnych prognoz inter- i ekstrapolacyjnych był podobny zarówno dla danych miesięcznych jak i dziennych.
 3. Wielkości ocen błędów prognoz interpolacyjnych determinowane były rozmieszczeniem luk, podczas, gdy błędów prognoz ekstrapolacyjnych ich liczbami. W przypadku szeregów miesięcznych wysokość błędów była uzależniona zarówno od odsetek jak i od rozmieszczenia luk. Wyższe błędy otrzymywano dla tych kombinacji, w których luki przypadły na miesiące, w których występowały minima lub maksima sezonowe.
 4. Zarówno dla szeregów o złożonej sezonowości jak i szeregów miesięcznych, zwłaszcza dla wariantów o dużej liczbie usuniętych podokresów, występowały kombinacje liniowe składowych, dla których niezbędne było oszacowanie kilku wersji modeli.
 5. Przedstawione w pracy wyniki, odnoszące się do wybranej zmiennej, charakteryzującej się dość silnym natężeniem każdego z rodzajów złożonych wahań sezonowych, mogą być uogólnione na inne zmienne o podobnym ich natężeniu. Wynika to z faktu, że rozpatrywanych było ponad pięć tysięcy możliwych układów luk o zróżnicowanych ich liczbach, odnoszących się zarówno do każdego z rodzajów wahań z osobna jak i obu rodzajów równocześnie.

BIBLIOGRAFIA

- Anderson D. R., Sweeney D. J., Williams T. A. (2011) *Statistics for business and economics* 11e. South-Western Cengage Learning.
- Oesterreich M. (2012) Symulacyjne badanie wpływu częstości występowania luk niesystematycznych w szeregach czasowych na dokładność prognoz. *Prace Naukowe UE we Wrocławiu, seria Ekonometria*, 38, 186–196.
- Oesterreich M. (2014a) Kombinacje liniowe składowych harmonicznych a dokładność prognoz w sezonowych szeregach czasowych z lukami systematycznymi. *Folia Pomeranae Universitatis Technologiae Stetinensis, seria Oeconomica* 313(76)3, 83–91.
- Oesterreich M. (2014b) Badanie wpływu częstości występowania luk systematycznych w mikroekonomicznym szeregu czasowym na dokładność prognoz. [w:] *Problemy społeczno-ekonomiczne w relacjach międzynarodowych: analiza modelowa rozwoju regionów*. Wydawnictwo UE w Katowicach, 219–230.

- Oesterreich M. (2017) Symulacyjne badanie wpływu liczby i rozmieszczenia luk na dokładność prognoz w szeregu czasowym dla danych dziennych. Prace Naukowe UE we Wrocławiu, seria: Ekonometria – artykuł w druku.
- Zawadzki J. (red.) (1999) Ekonometryczne metody predykcji dla danych sezonowych w warunkach braku pełnej informacji. Wydawnictwo Uniwersytetu Szczecińskiego.
- Zawadzki J. (red.) (2003) Zastosowanie hierarchicznych modeli szeregów czasowych w prognozowaniu zmiennych ekonomicznych z wahaniami sezonowymi. Wydawnictwo AR w Szczecinie.

**THE SIMULATION ANALYSIS OF THE IMPACT
OF THE SYSTEMATIC GAPS IN THE DAILY TIME SERIES
ON ACCURACY OF FORECASTS**

Abstract: In the paper was presented the simulation analysis of the impact of systematic gaps on the accuracy of inter- and extrapolative forecasts for daily time series. To forecasts construction were used classical time series model, in which a weekly and an annual seasonality was described by dummy variables. The analysed variable was daily sale of liquid fuels in liters in petrol station X in years 2012-2014. Data in years 2012-2013 were used in model construction and year 2014 was a period of empirical validation of forecasts. Eleven different variants of systematic gaps were examined. Calculations were made using the R statistical environment and the Statsoft Statistica12.

Keywords: high frequency time series, complex seasonal fluctuations, systematic gaps, forecasting, time series models