FOLIA POMERANAE UNIVERSITATIS TECHNOLOGIAE STETINENSIS
Folia Pomer. Univ. Technol. Stetin. 2010, Oeconomica 280 (59), 77–86

Maciej Oesterreich

WYKORZYSTANIE METOD NUMERYCZNYCH W PROGNOZOWANIU BRAKUJĄCYCH DANYCH W SZEREGACH CZASOWYCH Z SEZONOWOŚCIĄ

APPLICATION NUMERICAL METHODS IN PREDICTING MISSING DATA IN SEASONAL TIME SERIES

Katedra Zastosowań Matematyki w Ekonomii, Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
ul. Klemensa Janickiego 31, 71-270 Szczecin

Abstract. The following study presents the empirical analysis of the numeric methods in forecasting in conditions of lack of full information. In forecasting the following methods were used: segment, two variants of curves methods, and four variants of Lagrange methods. In analysis are used the average relative forecast errors in six variants of blanks. This study is an attempt to answer a question, whether the amount and distribution of blanks affect the quality of forecasts.

Słowa kluczowe: brakujące dane, metoda łuków, metoda odcinkowa, metody numeryczne.
Key words: curves method, missing data, numerical methods, segment method.

WSTĘP

Jednym z klasycznych warunków modelowania ekonometrycznego, zarówno w przypadku danych przekrojowych, jak i postaci szeregów czasowych, jest kompletność danych statystycznych. Występowanie luk wprawdzie utrudnia, ale nie uniemożliwia modelowania i prognozowania zmiennych ekonomicznych. Do prognozowania brakujących danych w szeregach czasowych mogą być wykorzystywane różne metody. Stosowanie poszczególnych metod wymaga spełnienia określonych założeń. Przykładowo w modelach adaptacyjnych szeregi czasowe powinny zawierać pewną liczbę obserwacji początkowych, które są niezbędne do wyznaczania wartości startowych. Z kolei inne metody numeryczne mogą być stosowane wyłącznie do obliczania prognoz interpolacyjnych. Ważnym czynnikiem wpływającym na wybór metody jest występowanie (lub niewystępowania) wahań sezonowych. Modele klasycznego szeregu czasowego, w których sezonowość opisywana jest ze pomocą zmiennych zero-jedynkowych, mogą służyć do modelowania i prognozowania w warunkach niepełnej informacji jedynie wówczas, gdy występują niesystematyczne luki w danych. Natomiast modele szeregu czasowego z wielomianem trygonometrycznym oraz modele hierarchiczne mogą być wykorzystywane zarówno w przypadku luk systematycznych, jak i niesystematycznych.

Istotnym ograniczeniem związanym ze stosowaniem metod numerycznych jest to, że mogą one być stosowane bezpośrednio w szeregach czasowych niezawierających wahań sezonowych lub takich, z których została wyeliminowana sezonowość (danych oczyszczonych). Procedura prognozowania przebiega następująco: W pierwszej kolejności eliminuje się sezonowość, a następnie oblicza prognozy tylko interpolacyjne lub inter- i ekstrapolacyjne. Tak otrzymane prognozy mnoży się przez wskaźniki sezonowości.

Tego rodzaju procedura zostanie wykorzystana w pracy w prognozowaniu skupu mleka. W przypadku niektórych metod numerycznych luki nie mogą występować na początku oraz na końcu szeregu (co niekiedy może być dużym ograniczeniem).

PRZEGLĄD METOD NUMERYCZNYCH WYKORZYSTANYCH W PRACY

Metoda odcinkowa

Metoda odcinkowa jest wykorzystywana do interpolacji brakujących danych statystycznych. Jest stosowana szczególnie wówczas, gdy „[…] liczba znanych wyrazów szeregu czasowego jest mała, a ich przebieg w czasie cechuje niewielka regularność, co uniemożliwia dopasowanie odpowiedniej funkcji trendu.

Istota tej metody polega na łączeniu znanych, sąsiadujących ze sobą wyrazów szeregu za pomocą odcinków prostej. Wartości nieznane elementów szeregu, które leżą między użytymi w procedurze wyrazami szeregu, określa się według wzoru:

$$y_0 = \frac{t_0 - t_2}{t_1 - t_2} y_1 + \frac{t_0 - t_1}{t_2 - t_1} y_2 \tag{1}$$

gdzie:

(y_1, t_1) oraz (y_2, t_2) – są współrzędne dwóch sąsiednich punktów, na podstawie których określa się równanie kolejnego odcinka.

Procedura ta może służyć wyłącznie do celów interpolacji, czyli szacowania brakujących wyrazów szeregu czasowego leżących między wyrazami znanymi. Jej wadą jest brak możliwości określenia *a priori* błędów dokonanych szacunków" (Dittmann 2003, s. 57).

Metoda łuków

Metoda łuków jest modyfikacją metody odcinkowej. Istota tej metody polega na łączeniu trzech znanych, sąsiadujących ze sobą, wyrazów szeregu za pomocą paraboli.

W metodzie łuków I interpolowane wartości szeregu czasowego y_1^* w okresie t_1 wyznacza się na podstawie równania paraboli przechodzącej przez trzy sąsiadujące ze sobą punkty: (y_0, t_0) na lewo od interpolowanego punktu oraz (y_2, t_2) i (y_3, t_3) na prawo od tego punktu:

$$y_1^* = \frac{(t_1 - t_2)(t_1 - t_3)}{(t_0 - t_2)(t_0 - t_3)} y_0 + \frac{(t_1 - t_0)(t_1 - t_3)}{(t_2 - t_0)(t_2 - t_3)} y_2 + \frac{(t_1 - t_0)(t_1 - t_2)}{(t_3 - t_0)(t_3 - t_2)} y_3 \tag{2}$$

Z kolei w metodzie łuków II interpolowaną wartość y_2^* w okresie t_2 wyznacza się na podstawie paraboli przechodzącej przez trzy punkty, z tą różnicą, że dwa z nich – (y_0, t_0) i (y_1, t_1) leżą na lewo od interpolowanego punktu, a (y_3, t_3) na prawo.

$$y_2^* = \frac{(t_2 - t_1)(t_2 - t_3)}{(t_0 - t_1)(t_0 - t_3)} y_0 + \frac{(t_2 - t_0)(t_2 - t_3)}{(t_1 - t_0)(t_1 - t_3)} y_1 + \frac{(t_2 - t_0)(t_2 - t_2)}{(t_3 - t_0)(t_3 - t_2)} y_3 \tag{3}$$

Obie procedury mogą służyć wyłącznie do celów interpolacji. Nie można ich stosować do ekstrapolacji. Podobnie jak w przypadku metody odcinkowej, metody te nie dają możliwości określenia *a priori* wielkości błędu interpolacji.

Metoda wielomianowa Lagrange'a

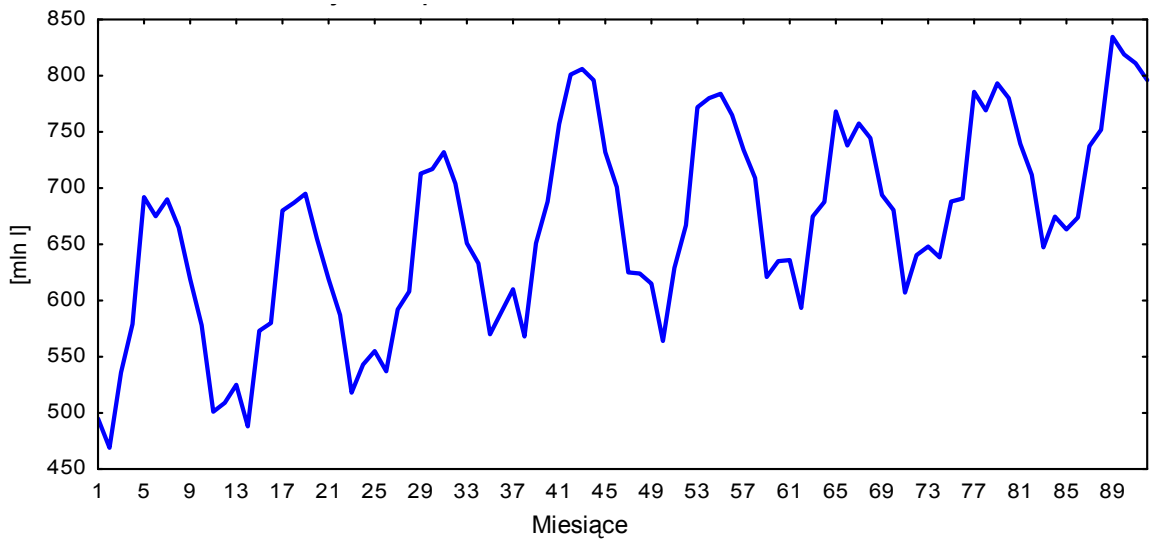
Metoda ta jest oparta na twierdzeniu, że jakkolwiek jest dana funkcja $f(x)$ i jakkolwiek są wybrane węzły interpolacji $x_0, x_1,..., x_n$, „[...] istnieje dokładnie jeden wielomian interpolacyjny $\varphi(x_i)$ stopnia n, który w punktach $x_0, x_1,..., x_n$, przyjmuje te same wartości, co dana funkcja $f(x)$" (Fortuna 2001, s. 25).

Istotą tej metody jest „[…] wyznaczenie przybliżonych wartości funkcji interpolacyjnej $f(t)$ w punktach, które nie są węzłami oraz oszacowanie błędów tych przybliżonych wartości. Funkcja ta w węzłach interpolacji przyjmuje takie same wartości co funkcja $y = f(t)$. Funkcji interpolacyjnej poszukuje się najczęściej jako funkcji pewnej określonej postaci. Zazwyczaj uznaje się, że funkcja $f(t)$ jest wielomianem potęgowym stopnia nie większego od k zmiennej t, co prowadzi do wykorzystania mającej największe zastosowanie praktyczne tzw. interpolacji parabolicznej. […] Wybór postaci funkcji interpolującej wpływa na ilość uzyskanych w wyniku jej zastosowania rozwiązań, których może być nieskończenie wiele lub może ich nie być wcale. Jest rzeczą bardzo ważną, aby istniała dokładnie jedna funkcja interpolująca" (Stoer 2001, s. 39). Wyróżnia się zazwyczaj dwa warianty tej metody różniące się sposobem rozmieszczenia węzłów. Mogą być one rozmieszczone proporcjonalnie albo zgodnie z procedurą optymalizacji Czebyszewa.

CHARAKTERYSTYKA ZMIENNEJ ORAZ ZAKRES BADAŃ

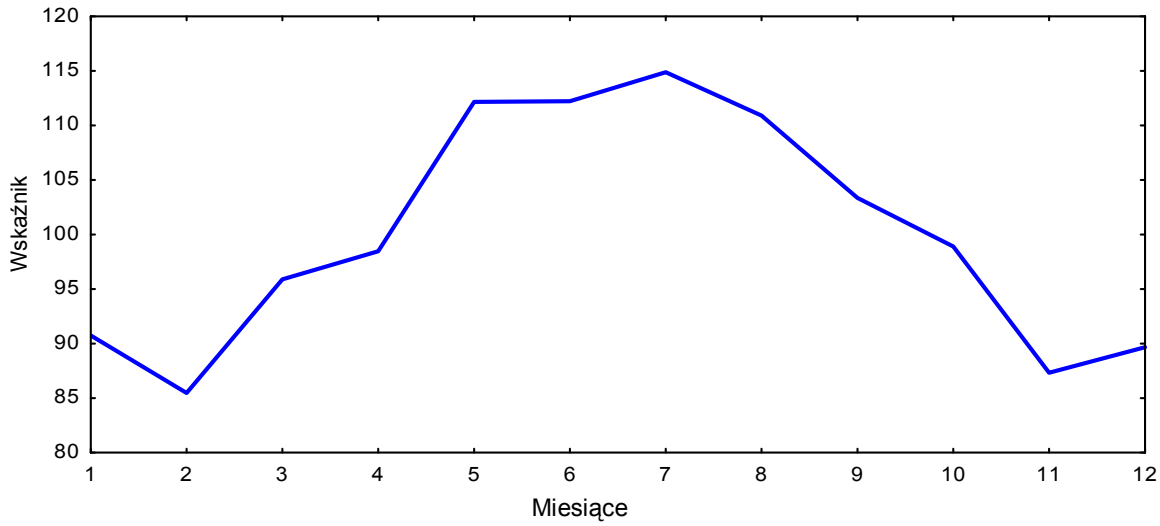
Analiza efektywności prognoz wyznaczonych za pomocą metod numerycznych zostanie przeprowadzona ma przykładzie kształtowania się wielkości skupu mleka w Polsce. Dane te pochodzą z miesięcznych Biuletynów Statystycznych Głównego Urzędu Statystycznego z okresu od stycznia 2002 roku do sierpnia 2009 roku. Lata 2002–2008 były okresem estymacyjnym. Natomiast dane z ośmiu miesięcy roku 2009 zostały wykorzystane w procesie empirycznej weryfikacji prognoz.

Kształtowanie się wielkości skupu mleka w zostało przedstawione na rys. 1. Wynika z niego, że zmienna charakteryzuje się występowaniem silnych wahań sezonowych.



Rys. 1. Skup mleka w Polsce w latach 2002–2009
Źródło: opracowanie własne na podstawie Biuletynów Statystycznych GUS z lat 2002–2009.

W celu określenia natężenia tych wahań obliczono wskaźniki sezonowe. Oceny tych wskaźników przedstawiono na rys. 2.



Rys. 2. Oceny wskaźników sezonowych dla skupu mleka

Maksymalnymi ocenami wskaźników charakteryzują się miesiące letnie, a minimalnymi miesiące zimowe. Różnica pomiędzy okresami o najwyższym i najniższym natężaniu skupu mleka wynosi 29 punktów procentowych.

Jak wspomniano wyżej, metody numeryczne nie mogą być wykorzystane bezpośrednio do budowy prognoz z wahaniami sezonowymi. Dlatego przed ich zastosowaniem z oryginalnego szeregu wyeliminowano sezonowość, dzieląc jego wartości przez wskaźniki sezonowe.

Rozpatrywanych było sześć niżej wymienionych wariantów luk:

– wariant I – luki w miesiącach nieparzystych,

– wariant II – luki w I i III kwartale każdego roku,

– wariant III – luki w II i IV kwartale każdego roku,

– wariant IV – luki w I i II miesiącu każdego kwartału,

– wariant V – luki w I i III miesiącu każdego kwartału,

– wariant VI – luki w II i III miesiącu każdego kwartału.

Warianty te otrzymano poprzez „wymazanie" części danych z oryginalnego szeregu.

Analiza dokładności prognoz będzie polegała na wyznaczeniu średnich względnych różnic między prognozami inter- i ekstrapolacyjnymi a realizacjami badanej zmiennej. Dla prognoz ekstrapolacyjnych została przeprowadzona analiza *ex*-post ich dokładności. Prognozy ekstrapolacyjne wyznaczono dla horyzontu h = 8.

Do budowy prognoz wykorzystano następujące metody (w nawiasach podano odpowiadające im skróty):

– odcinkową (O),

– łuków I (Ł1),

– łuków II (Ł2),

– Lagrange'a z trzema węzłami rozmieszczonymi proporcjonalnie (L-3 WP),

– Lagrange'a z trzema węzłami rozmieszczonymi zgodnie z funkcją optymalizacji Czebyszewa (L-3 WC),

– Lagrange'a z czteroma węzłami rozmieszczonymi proporcjonalnie (L-4 WP),

– Lagrange'a z czteroma węzłami rozmieszczonymi zgodnie z funkcją optymalizacji Czebyszewa (L-4 WC).

Ze względu na specyfikę metody odcinkowej oraz łuków oceny błędów dla prognoz ekstrapolacyjnych przeprowadzono metodą najmniejszych kwadratów (MNK). Oszacowano je, wykorzystując trend wielomianowy ze zmienną stopą wzrostu.

Dla metody Lagrange'a węzły rozłożone proporcjonalnie zostały „zakotwiczone" w następujący sposób:

– dla trzech węzłów w 1, 42 oraz 84 obserwacji,

– dla czterech węzłów w 1, 28, 56 oraz 84 obserwacji.

W drugim wariancie tej metody trzy lub cztery węzły zostały rozmieszczone według procedury optymalizacji Czebyszewa.

WYNIKI PROGNOZOWANIA INTERPOLACYJNEGO

W tabeli 1 zamieszczono oceny średnich względnych błędów prognoz interpolacyjnych skupu mleka dla poszczególnych wariantów luk danych otrzymanych siedmioma metodami.

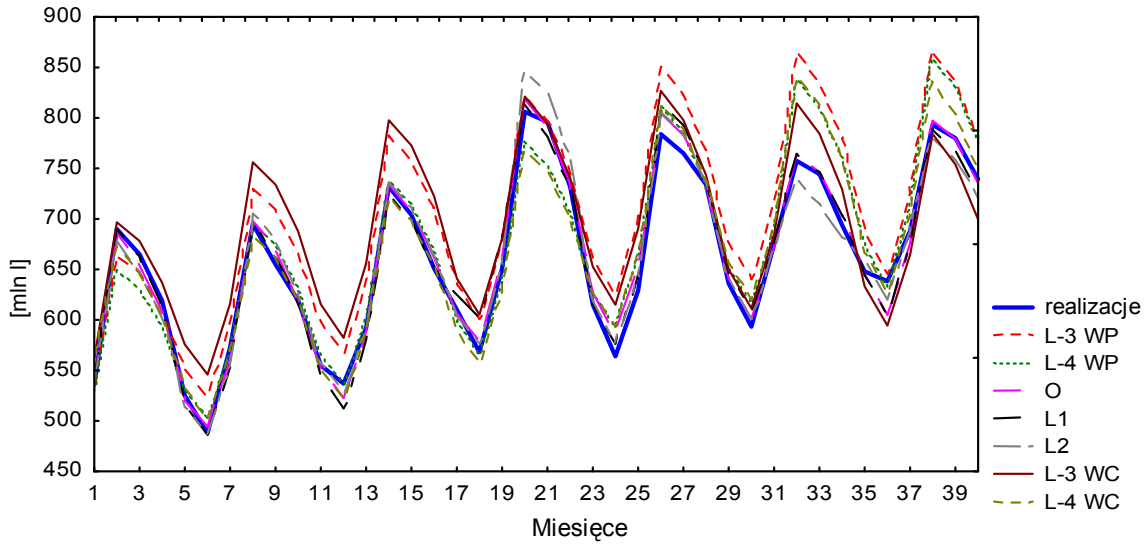
Tabela 1. Oceny średnich względnych prognoz interpolacyjnych [%]

Wariant	Metoda	Błąd	Metoda	Błąd
I	O	1,16	L-3 WP	6,30
	Ł1	1,21	L-3 WC	5,46
	Ł2	1,21	L-4 WP	3,32
			L-4 WC	2,70
II	O	1,35	L-3 WP	6,34
	Ł1	1,76	L-3 WC	5,85
	Ł2	1,85	L-4 WP	3,45
			L-4 WC	3,25
III	O	1,60	L-3 WP	6,16
	Ł1	1,69	L-3 WC	5,11
	Ł2	2,19	L-4 WP	3,38
			L-4 WC	2,90
IV	O	1,22	L-3 WP	6,36
	Ł1	1,26	L-3 WC	5,53
	Ł2	1,39	L-4 WP	3,43
			L-4 WC	2,95
V	O	1,32	L-3 WP	6,16
	Ł1	1,54	L-3 WC	5,11
	Ł2	1,41	L-4 WP	3,43
			L-4 WC	3,13
VI	O	1,41	L-3 WP	6,24
	Ł1	1,44	L-3 WC	4,53
	Ł2	1,37	L-4 WP	3,38
			L-4 WC	3,26

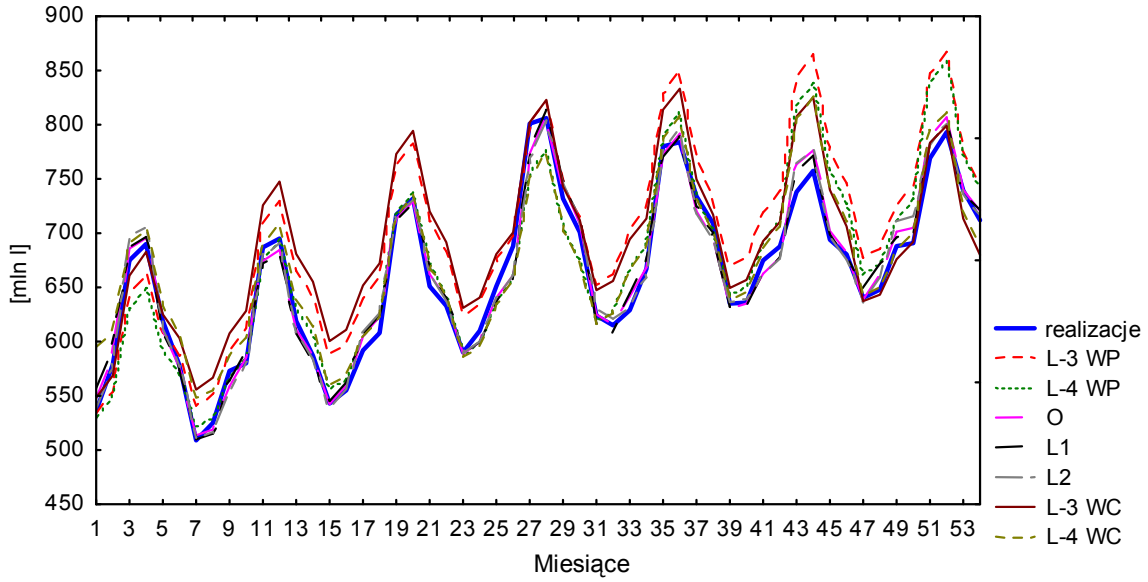
Z informacji zawartych w tab. 1 wynika, że błędy prognoz interpolacyjnych kształtują się w zakresie 1,16–6,36%. Najniższe oceny błędów dla poszczególnych wariantów nie przekraczają 1,6%. W przypadku pierwszych pięciu wariantów luk w danych minimalnymi ocenami błędów charakteryzują się prognozy otrzymane za pomocą metody odcinkowej.

Przyjmują one wartości z przedziału od 1,16% (wariant I) do 1,6% (wariant III). Jedynie w przypadku wariantu VI nieznacznie niższą ocenę otrzymano dla metody łuków II. Wśród prognoz otrzymanych metodą Lagrange'a zdecydowanie dokładniejsze są prognozy uzyskane dla czterech węzłów interpolacyjnych. W przypadku wszystkich wariantów nieco niższe oceny otrzymano dla węzłów rozmieszczonych według reguły Czebyszewa. Ich oceny błędów zawarte były w przedziale od 2,70% (wariant I) do 3,38% (wariant VI).

Najmniej efektywna okazała się metoda Lagrange'a dla trzech węzłów rozmieszczonych proporcjonalnie. W przypadku wszystkich wariantów oceny błędów przekraczały 6%. W przypadku wariantów I–V dla trzech węzłów rozmieszczonych zgodnie z regułą optymalizacji Czebyszewa błędy prognoz zawarte były w przedziale od 5,11% (wariant V) do 5,85% (wariant II). Jedynie w wariancie VI ocena błędu była niższa – wyniosła 4,53%. Oznacza to, że rozmieszczenie luk wpływa na dokładność prognoz. Kształtowanie się prognoz interpolacyjnych dla dwóch wariantów (II i V) przedstawiono w formie graficznej na rys. 3 i 4.



Rys. 3. Prognozy interpolacyjne skupu mleka – wariant II



Rys. 4. Prognozy interpolacyjne skupu mleka – wariant V

ANALIZA DOKŁADNOŚCI PROGNOZ EKSTRAPOLACYJNYCH

W tabeli 2 zamieszczone zostały oceny błędów prognoz obliczone dla okresu empirycznej weryfikacji prognoz obejmującego miesiące od stycznia do sierpnia 2009 roku.

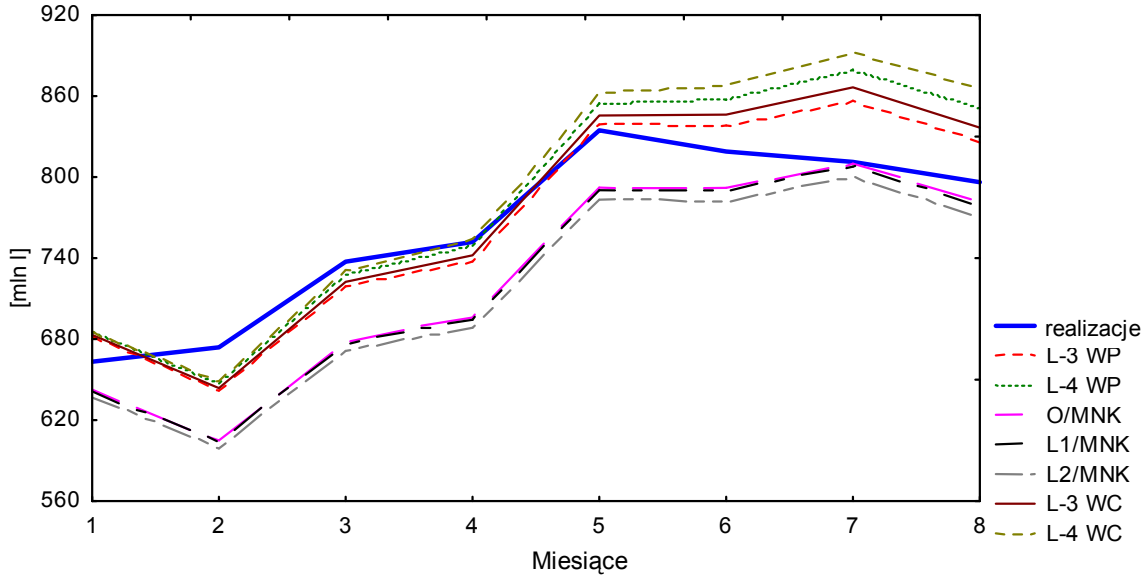
Tabela 2. Oceny średnich względnych prognoz ekstrapolacyjnych [%]

Wariant	Metoda	Błąd	Metoda	Błąd
I	O	4,39	L-3 WP	3,02
	Ł1	4,44	L-3 WC	3,42
	Ł2	4,43	L-4 WP	3,95
			L-4 WC	4,22
II	O	4,9	L-3 WP	3,02
	Ł1	5,13	L-3 WC	3,42
	Ł2	5,96	L-4 WP	3,95
			L-4 WC	4,56
III	O	4,24	L-3 WP	3,02
	Ł1	4,96	L-3 WC	3,16
	Ł2	4,47	L-4 WP	3,95
			L-4 WC	4,07
IV	O	4,64	L-3 WP	3,02
	Ł1	4,68	L-3 WC	3,73
	Ł2	4,71	L-4 WP	3,95
			L-4 WC	4,32
V	O	3,61	L-3 WP	3,02
	Ł1	3,61	L-3 WC	3,13
	Ł2	3,68	L-4 WP	3,95
			L-4 WC	5,32
VI	O	4,78	L-3 WP	3,02
	Ł1	4,97	L-3 WC	3,40
	Ł2	4,83	L-4 WP	3,95
			L-4 WC	4,22

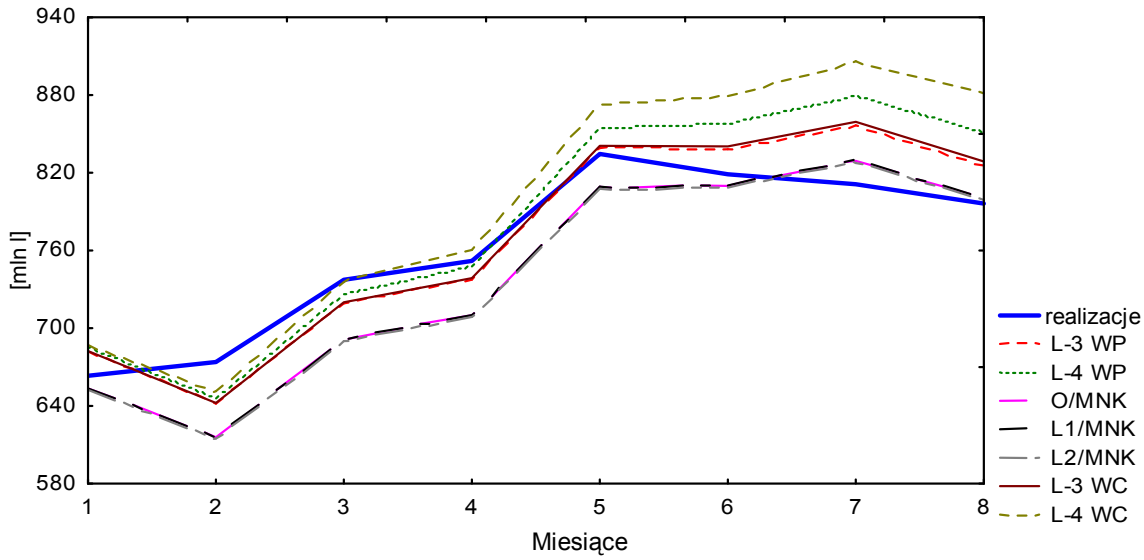
Prognozy ekstrapolacyjne dla metod odcinkowej oraz łuków I i łuków II otrzymano w sposób pośredni. Polegało to na tym, że szeregi czasowe z lukami w danych uzupełniono o prognozy interpolacyjne, a następnie za pomocą MNK zbudowano modele trendów o zmiennej stopie wzrostu. Na ich podstawie wyznaczono prognozy ekstrapolacyjne. W przypadku metody Lagarange'a prognozy zbudowano bezpośrednio. W obu przypadkach były to prognozy dla danych oczyszczonych. Prognozy ostateczne otrzymano po ich przemożeniu przez wskaźniki sezonowości.

Z informacji zawartych w tab. 2 wynika, że najniższymi ocenami błędów charakteryzowały się prognozy otrzymane za pomocą metody Lagrange'a dla trzech węzłów rozmieszczonych proporcjonalnie. Metoda ta, na podstawie przeprowadzonej wcześniej analizy, okazała się najmniej efektywna w prognozowaniu interpolacyjnym. Drugą w kolejności była metoda Lagrange'a dla trzech węzłów rozmieszczonych zgodnie z reguła Czebyszewa. Błędy prognoz, otrzymanych metodami odcinkową oraz łuków I i łuków II dla wariantu I–IV oraz VI, są od 1,22 do 2,94 punktu procentowego wyższe od błędu minimalnego. Jedynie w przypadku

wariantu V różnica wynosi 0,66%. Oceny błędów prognoz otrzymanych tymi metodami wykazują jednocześnie duże zróżnicowanie. Różnica ocen błędów minimalnych między wariantami skrajnym II i V wynosi 2,35 punktu procentowego. Najwyższymi ocenami błędów cechują się metody łuków I i łuków II. Rzeczywiste kształtowanie się skupu mleka oraz prognozy ekstrapolacyjne dla pierwszych 8 miesięcy 2009 roku dla wariantów II i V przedstawiono graficznie na rys. 5 i 6.



Rys. 5. Prognozy interpolacyjne skupu mleka – wariant II



Rys. 6. Prognozy interpolacyjne skupu mleka – wariant V

PODSUMOWANIE

Z przeprowadzonych w pracy rozważań wynika, że wykorzystanie wybranych metod numerycznych w prognozowaniu brakujących danych okazało się efektywne.

Należy jednak zauważyć, że inne metody powinny być wykorzystane w prognozowaniu interpolacyjnym, a inne w prognozowaniu ekstrapolacyjnym. Prognozy interpolacyjne o najniższych błędach otrzymano metodą odcinkową. W przypadku prognoz ekstrapolacyjnych była to metoda Lagrange'a z trzema węzłami rozmieszczonymi proporcjonalnie, a następnie ta sama metoda dla trzech węzłów rozmieszczonych zgodnie z regułą Czebyszewa. W toku badań stwierdzono, że liczba oraz rozmieszczanie luk wpływają na dokładność prognoz zarówno inter-, jak i ekstrapolacyjnych.

PIŚMIENNICTWO

Dittmann P. 2003. Prognozowanie w przedsiębiorstwie. Metody i ich zastosowanie. Kraków, Oficyna Ekonomiczna.

Fortuna Z. 2001. Metody numeryczne. Warszawa, Wydawnictwo Naukowo-Techniczne.

Stoer J., Bulirsch R. 2001. Introduction to numerical analysis. New York, Wyd. Springer-Verlag.

Ralston A. 1965. Wstęp do analizy numerycznej. Warszawa, PWN.

Zboś D. 1995. Metody numeryczne. Kraków, Politechnika Krakowska.