

Alicja Markiewicz, Jan Zawadzki

Akademia Rolnicza w Szczecinie

EMPIRYCZNA ANALIZA PORÓWNAWCZA DOKŁADNOŚCI WYBRANYCH METOD PROGNOZOWANIA W WARUNKACH BRAKU PEŁNEJ INFORMACJI

1. Wstęp

Artykuł jest poświęcony badaniu efektywności klasycznych modeli szeregu czasowego, modeli hierarchicznych oraz modeli sieci neuronowych w prognozowaniu brakujących danych zmiennych charakteryzujących się różnym natężeniem sezonowości.

Modelowanie i prognozowanie dla danych w postaci szeregów czasowych w sytuacji, gdy zachodzą szybkie zmiany o charakterze strukturalnym, jest dość skomplikowane, ponieważ należy wprowadzić do modelu „mechanizm” opisujący te zmiany. Ze względu jednak na to, że zasadniczy wpływ na nowo budowane prognozy mają informacje najnowsze lub też pochodzące z niezbyt odległej przeszłości, zagadnienie opisu zmian strukturalnych można do pewnego stopnia „obejść”, odrzucając informacje najstarsze.

Wykorzystanie klasycznych modeli szeregu czasowego z wahaniami sezonowymi (ze zmiennymi zero-jedynkowymi lub wielomianem trygonometrycznym) do szacowania parametrów w sposób istotny zmniejsza liczbę stopni swobody. W przypadku stałych bądź relatywnie stałych wahań sezonowych szacuje się $m - 1$ parametrów. Dla zmiennych wahań liczba ta ulega podwojeniu. Wyjściem z tej sytuacji może być wykorzystanie modeli hierarchicznych.

Modele hierarchiczne mają tę ważną zaletę, że umożliwiają prognozowanie na podstawie krótkich szeregów czasowych. Maksymalna liczba szacowanych parametrów w modelach opisujących sezonowość periodyczną równa jest połowie długości cyklu wahań. Liczba uporządkowanych hierarchicznych modeli szeregu czasowego zależy od długości m cyklu wahań okresowych i jest równa liczbie permutacji i permutacji z powtórzeniami dzielników p , długości tego okresu. W każdym uporządkowanym modelu

hierarchicznym podzielniki spełniają jednocześnie dwa warunki (zob. [Zastosowanie hierarchicznych modeli... 2003]):

$$2 \leq p_i \leq \frac{m}{2} \quad (i = 1, \dots, r), \quad (1)$$

$$\prod_{i=1}^r p_i = m. \quad (2)$$

W przypadku danych miesięcznych wspomniana liczba kombinacji wynosi siedem. Wśród siedmiu modeli hierarchicznych są cztery modele dwustopniowe i trzy trójstopniowe.

Niżej podamy przykładowy zapis dwupoziomowego hierarchicznego modelu szeregu czasowego z liniowym trendem i stałą sezonowością:

$$Y_{srt} = \alpha_1 t + \alpha_0 + \sum_{s=1}^{\frac{m}{p_1}} b_{0s} Q_{st} + \sum_{r=1}^{\frac{m}{p_1 p_2}} b_{0sr} Q_{srt} + U_{srt}, \quad (3)$$

gdzie:

$$\sum_{s=1}^{\frac{m}{p_1}} b_{0s} = \sum_{r=1}^{\frac{m}{p_1 p_2}} b_{0sr} = 0. \quad (4)$$

2. Opis metod prognozowania brakujących danych

Do budowy prognoz brakujących danych zostaną wykorzystane modele hierarchiczne, klasyczne oraz modele sieci neuronowych. Modele hierarchiczne mogą być wykorzystywane zarówno dla luk niesystematycznych, jak i systematycznych.

Kolejną z metod prognozowania brakujących danych w szeregach czasowych jest metoda oparta na koncepcji budowy modeli „oszczędnych”. Polega ona na uwzględnianiu istotnych statystycznie składowych sinuso- i cosinusoidalnych w modelach, w których stała sezonowość opisywana jest za pomocą wielomianu trygonometrycznego:

$$Y_t = \alpha_1 t + \alpha_0 + \sum_k^{\frac{m}{2}} (\alpha_{0k} \sin \omega_i t + b_{0k} \cos \omega_i t), \quad (5)$$

gdzie: $\omega_i = \frac{2\pi \cdot i}{m}$.

W ostatnich latach podejmowane są także dość liczne próby wykorzystania w prognozowaniu zmiennych ekonomicznych sieci neuronowych. Sieci te stanowią wielowymiarową, nieliniową oraz nieparametryczną technikę wnioskowania, która nie jest zależna od modelu. Posiadają charakter wielowymiarowy, ponieważ wejście sieci neuronowej mogą zawierać dużą liczbę zmiennych. Wzajemne relacje oraz zależności przyczynowo-skutkowe są wykorzystane do prognozowania szeregu w przyszłości. Nieparametryczność i niezależność od modelu jest konsekwencją braku założeń dotyczących relacji zarówno między zmiennymi wejściowymi, jak i wielkościami ekstrapolowanymi. Do specyfikacji modelu odwzorowania danych wejściowych na zbiór danych wyjściowych nie są wymagane żadne z góry zadane parametry. Należy zwrócić uwagę, że liczba szacowanych parametrów sieci neuronowej rośnie liniowo do osiągnięcia tego samego stopnia aproksymacji [Dintcheva-Biś 1999].

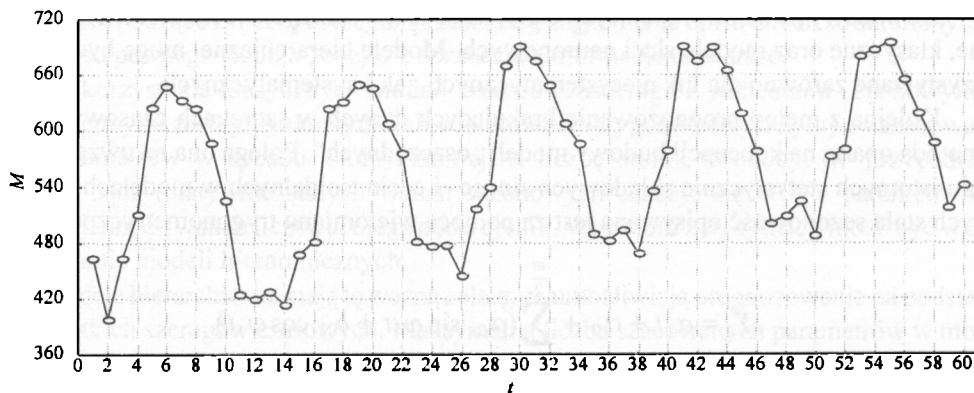
3. Przykład empiryczny

Analizie porównawczej dokładności prognoz za pomocą trzech przedstawionych wyżej metod poddano trzy zmienne charakteryzujące się różnym natężeniem wahań sezonowych:

- M – skup mleka,
- WK – wydobywanie węgla kamiennego,
- TUR – liczba osób korzystających z noclegów w obiektach turystycznych.

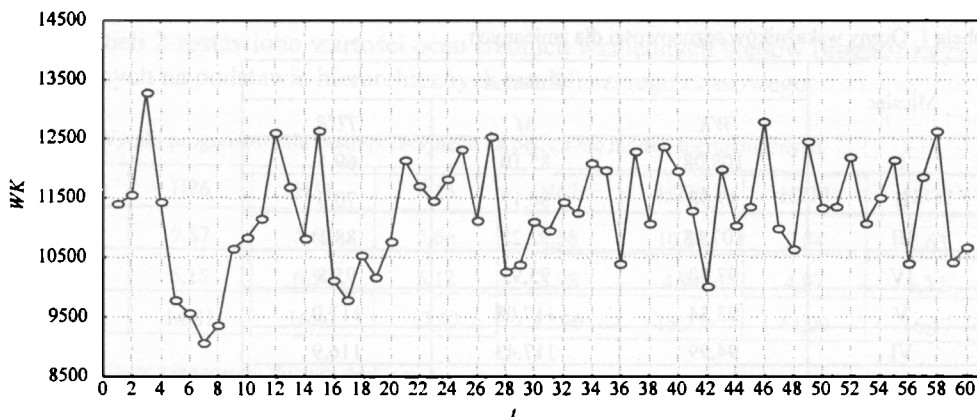
Przedział czasowy „próby” dla zmiennych o umiarkowanym natężeniu sezonowości M i WK obejmował 48 miesięcy, natomiast dla zmiennej charakteryzującej się silnym natężeniem sezonowości (TUR) obejmował 96 miesięcy. Dla następnych dwunastu miesięcy budowane były prognozy ekstrapolacyjne. Ponieważ znane były dla nich realizacje, przeprowadzono analizę *ex post* ich dokładności.

Kształtowanie się badanych zmiennych przedstawiono na rys. 1-3.

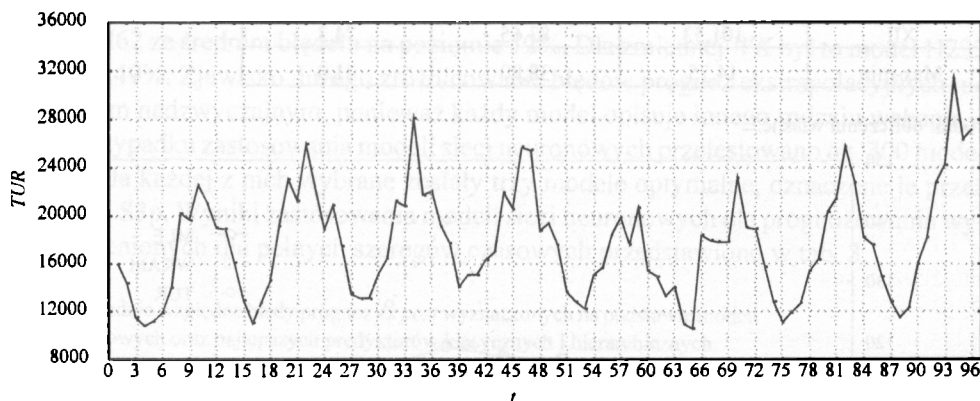


Rys. 1. Kształtowanie się zmiennej M

Źródło: opracowanie własne.

Rys. 2. Kształtowanie się zmiennej WK

Źródło: opracowanie własne.

Rys. 3. Kształtowanie się zmiennej TUR

Źródło: opracowanie własne.

W tabeli 1 przedstawiono kształtowanie się wartości ocen wskaźników sezonowości dla poszczególnych zmiennych.

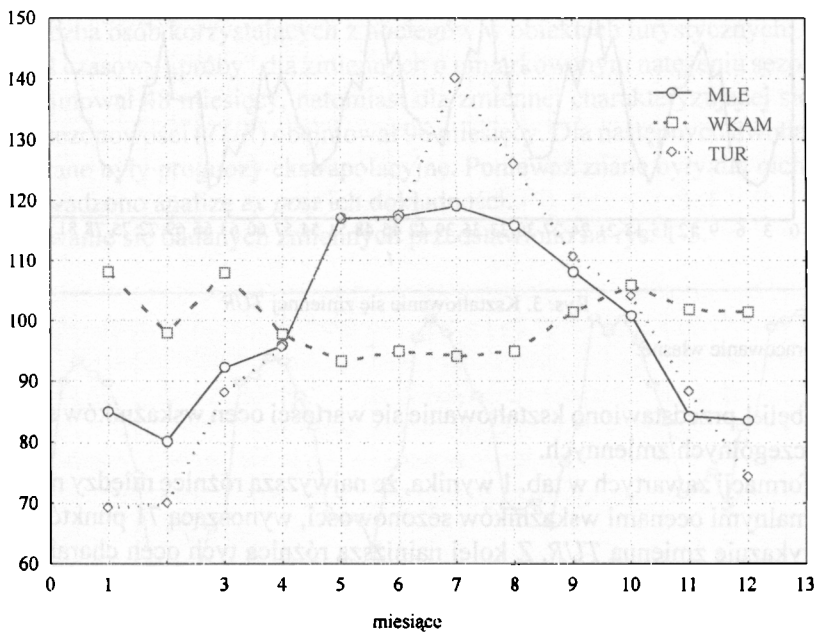
Z informacji zawartych w tab. 1 wynika, że najwyższą różnicę między minimalnymi i maksymalnymi ocenami wskaźników sezonowości, wynoszącą 71 punktów procentowych, wykazuje zmienna TUR . Z kolei najniższą różnicą tych ocen charakteryzuje się zmienna WK . Maksima sezonowe zmiennych M i TUR przypadają na miesiące letnie (czerwiec, lipiec, sierpień), natomiast dla zmiennej WK maksimum to przypada w styczniu.

Kształtowanie się wskaźników sezonowości w formie graficznej przedstawiono na rys. 4.

Tabela 1. Oceny wskaźników sezonowości dla zmiennych

Miesiąc	Zmienna		
	<i>WK</i>	<i>M</i>	<i>TUR</i>
I	108,08	85,10	69,1
II	97,98	80,17	70,0
III	107,88	92,28	88,1
IV	97,80	95,93	95,5
V	93,34	117,08	117,0
VI	94,99	117,43	116,9
VII	94,04	119,07	140,1
VIII	95,00	115,95	126,1
IX	101,48	108,25	110,6
X	106,02	100,89	104,0
XI	101,85	84,22	88,3
XII	101,53	83,65	74,3
Max-min	14,75	38,90	71,0

Źródło: obliczenia własne.



Rys. 4. Wskaźniki sezonowości

Źródło: opracowanie własne.

W tabeli 2 zestawiono wartości ocen średnich względnych błędów prognoz *ex post* otrzymanych na podstawie hierarchicznych modeli szeregu czasowego.

Tabela 2. Wyniki prognozowania ekstrapolacyjnego na podstawie modeli hierarchicznych

Zmienna	H26	H34	H43	H62	H223	H232	H322
<i>M</i>	9,87	5,67	5,66	3,58	10,20	9,84	5,63
<i>WK</i>	5,15	6,50	6,12	5,98	4,94	4,97	6,11
<i>TUR</i>	14,81	16,17	12,92	12,00	15,27	14,09	16,17

Oceny najniższe zaznaczono tłustym drukiem.

Źródło: obliczenia własne.

Najniższy średni błąd prognozy (2,62%) wyznaczony na podstawie modelu H62 uzyskano dla zmiennej *M*. W przypadku zmiennej o silnym natężeniu sezonowości *TUR*, modelem o najlepszych własnościach predyktywnych okazał się również model hierarchiczny H62 ze średnim błędem na poziomie 12%. Dla zmiennej *WK* był to model H232 z oceną 4,49%. Zjawisko dużego zróżnicowania błędów prognoz ekstrapolacyjnych nie jest niczym nadzwyczajnym, ponieważ każdy model opisuje innego rodzaju wahania.

W przypadku zastosowania modeli sieci neuronowych przetestowano ok. 300 modeli sieci. Dla każdej z nich wybrane zostały trzy modele optymalne, oznaczone je przez S1p, S2p i S3p. Wyniki zastosowania modeli sieci neuronowych dla prognozowania wyżej wymienionych dla pełnych szeregów czasowych przedstawiono w tab. 3.

Tabela 3. Średnie względne błędy prognoz *ex post* wyznaczonych na podstawie modeli sieci neuronowych oraz najlepszych predyktorów klasycznych i hierarchicznych.

Zmienna	Model	Neur.	Hier.	Tryg.
<i>M</i>	S1p	3,18	3,58	3,27
	S2p	3,40		
	S3p	3,20		
<i>WK</i>	S1p	5,55		
	S2p	5,48	4,94	6,46
	S3p	5,86		
<i>TUR</i>	S1p	14,20		
	S2p	18,73	12,00	10,89
	S3p	20,01		

Źródło: obliczenia własne.

Z informacji w niej podanych wynika, że dla zmiennych *M* i *TUR* najlepszy okazał się model S1p, a dla zmiennej *WK* model S2p. W tabeli zestawiono także minimalne oce-

ny błędów prognoz *ex post* otrzymanych na podstawie modeli hierarchicznych (kolumna Hier.) oraz modeli oszczędnych (kolumna Tryg.).

Z porównania trzech klas modeli wynika, że tylko dla zmiennej *M* błąd względny otrzymany dla sieci neuronowej był niższy od błędu dla prognoz hierarchicznych. Natomiast dla modelu oszczędnego jedynie dla zmiennej *TUR* prognozy były dokładniejsze od wyznaczonych dla sieci neuronowych.

Dotąd zakładano, że szeregi czasowe nie zawierają luk w danych. Jednak zgromadzenie kompletnych danych statystycznych, będące jednym z podstawowych wymogów statystyczno-ekonometrycznego badania zjawisk gospodarczych, z wielu względów może być bardzo trudne bądź wręcz niewykonalne. Dlatego też w pracy podjęto próbę prognozowania na podstawie szeregów z systematycznymi lukami w danych. Rozpatrywane były dwa warianty luk:

A – luki występujące w drugim i ósmym miesiącu cyklu rocznego,

B – luki występujące w co drugim miesiącu i obejmujące miesiące nieparzyste.

Liczba oraz rozmieszczenie luk w danych było determinowane przede wszystkim występowaniem maksimów i minimów sezonowych. Otrzymano je poprzez usunięcie odpowiednich danych z pełnych szeregów. Dla szeregów z lukami wyznacza się także prognozy interpolacyjne dla okresów, w których dane nie są dostępne.

Modelowanie predyktywne na podstawie szeregów czasowych z lukami systematycznymi zarówno dla modeli hierarchicznych, jak i dla modeli oszczędnych wiąże się z daleko idącymi konsekwencjami. W obu typach modeli składowe należące odpowiednio do różnych stopni w modelach hierarchicznych oraz różnych składowych sinuso- i cosinusoidalnych mogą być: skorelowane, współliniowe bądź stałe. Niektóre składowe mogą być kombinacjami liniowymi innych składowych. Wystąpienie kombinacji liniowych oraz współliniowości oznacza, że będzie szacowanych wiele wersji tego samego modelu. Wersje te charakteryzują się identycznymi ocenami: współczynników determinacji, odchyłeń standardowych, składników losowych oraz statystyk Durбина-Watsona. Ze statystycznego punktu widzenia są one więc nierozróżnialne. Dopiero analiza dokładności prognoz inter- i ekstrapolacyjnych pozwala wybrać wersje najlepsze.

Tabela 4. Średnie względne błędy prognoz interpolacyjnych otrzymanych na podstawie modeli hierarchicznych

Zmienna	Wariant	H26	H34	H43	H62	H223	H232	H322
<i>M</i>	A	19,25	7,29	11,94	5,22	22,28	22,27	11,39
	B	13,32	6,83	5,83	4,49	13,69	13,84	5,80
<i>WK</i>	A	5,00	6,53	9,22	5,95	5,72	3,82	6,32
	B	8,89	7,35	6,85	6,83	8,56	7,75	6,87
<i>TUR</i>	A	10,46	28,68	11,31	18,01	12,47	9,28	21,82
	B	15,51	17,20	13,90	17,31	15,71	13,27	17,22

Źródło: obliczenia własne.

W tabeli 4 przedstawiono wyniki prognozowania interpolacyjnego w przypadku braku pełnej informacji, prognoz otrzymanych na podstawie modeli hierarchicznych.

Z tabeli tej wynika, że w trzech przypadkach najlepsze okazały się predyktory oparte na modelu H62 i w trzech na modelu H232. Dla zmiennych M i TUR były to te same predyktory dla obu rodzajów luk. Natomiast dla zmiennej WK były różne. W wariancie A był to model H62, a w wariancie B model H232. Z porównania ocen minimalnych wynika, że dla zmiennych WK i TUR zwiększenie liczby luk z dwóch (wariant A) do sześciu (wariant B) wpłynęło na zwiększenie błędów prognoz interpolacyjnych. W przypadku zmiennej M było odwrotnie, błąd względny dla wariantu B był o blisko jeden punkt procentowy niższy niż dla wariantu A.

W tab. 5, podobnie jak dla pełnych danych, zamieszczone zostały błędy prognoz interpolacyjnych otrzymanych na podstawie modeli sieci neuronowych (Neur.), najlepszych modeli hierarchicznych (Hier.) oraz modeli oszczędnych (Tryg.).

Tabela 5. Średnie względne błędy prognoz interpolacyjnych dla wariantów A i B

Zmienna	Wariant	Sieć	Neur	Hier.	Tryg.
M	A	S1p	7,92		
		S2p	8,53		
		S3p	7,21	5,22	3,78
	B	S1p	16,22		
		S2p	12,49	4,49	7,57
		S3p	21,07		
WK	A	S1p	2,08	3,82	12,93
		S2p	6,74		
		S3p	9,55		
	B	S1p	5,89	6,83	8,88
		S2p	5,92		
		S3p	7,97		
TUR	A	S1p	13,37		
		S2p	14,24		
		S3p	12,52	9,28	10,45
	B	S1p	23,18		
		S2p	16,94	13,27	13,81
		S3p	26,76		

Źródło: obliczenia własne.

W przypadku sieci neuronowych dla zmiennej WK najlepszy dla obu rodzajów luk okazał się model S1p, a dla zmiennej TUR model S2p. Natomiast dla zmiennej M dla wa-

riantu A był to model S3p, a dla wariantu B model S2p. Widoczny jest wpływ liczby luk na dokładność prognoz. Dla wariantu B są one przynajmniej o trzy punkty procentowe wyższe. Błędy prognoz hierarchicznych są w pięciu na sześć przypadków niższe niż otrzymane na podstawie sieci neuronowej (z wyjątkiem wariantu B zmiennej *WK*). Modele oszczędne są efektywniejsze od sieci neuronowych dla zmiennych *M* i *TUR*. Porównanie modeli hierarchicznych i modeli oszczędnych w każdym przypadku wypada na korzyść modeli hierarchicznych. Zwiększenie liczby luk w danych sprawiło, że błędy prognoz dla wariantu B zmiennych *WK* i *TUR* były wyższe niż w wariantie A.

W tab. 6 zamieszczono oceny średnich względnych błędów prognoz ekstrapolacyjnych otrzymanych na podstawie predyktorów hierarchicznych.

Tabela 6. Wyniki prognozowania ekstrapolacyjnego na podstawie modeli hierarchicznych

Zmienna	Wariant	H26	H34	H43	H62	H223	H232	H322
<i>M</i>	A	10,22	5,95	6,36	3,84	10,74	10,60	6,33
	B	9,77	6,91	5,48	2,92	9,89	9,95	5,48
<i>WK</i>	A	5,72	7,48	6,43	5,97	5,60	5,01	6,07
	B	5,31	6,22	5,80	6,04	6,21	5,58	5,90
<i>TUR</i>	A	13,05	16,96	14,72	10,95	15,32	14,78	18,06
	B	12,22	17,04	12,38	10,92	12,95	16,04	20,70

Źródło: obliczenia własne.

Predyktorem najbardziej efektywnym czterokrotnie okazał się predyktor H62 i dwukrotnie predyktor H232. Dla obu wariantów luk za każdym razem są to te same predyktory. Zwraca uwagę fakt, że jedynie dla zmiennej *WK* błąd dla wariantu B był wyższy niż dla wariantu A (o 0,57 punktu procentowego). Dla pozostałych zmiennych był on bądź niemal identyczny (*TUR*), bądź nawet o blisko jeden punkt procentowy niższy (*M*).

Wyniki prognozowania ekstrapolacyjnego na podstawie sieci neuronowych, najlepszych wariantów modeli hierarchicznych oraz modeli oszczędnych dla obu wariantów luk w danych zestawione zostały w tab. 7.

Czterokrotnie najlepszym modelem neuronowym okazał się model S2p. Najniższą ocenę błędu prognoz otrzymano na podstawie modelu S1p dla wariantu B zmiennej *WK*, a na podstawie modelu S3p dla zmiennej *TUR*, także dla wariantu B. Błędy prognoz zmiennych *M* i *TUR* rosły wraz ze zwiększaniem się liczby luk.

Tabela 7. Średnie względne błędy prognoz ekstrapolacyjnych dla wariantów A i B

Zmienna	Wariant	Sieć	Neur.	Hier.	Tryg.
M	A	S1p	7,35		
		S2p	4,67	3,48	3,95
		S3p	26,91		
	B	S1p	9,19		
		S2p	6,71	2,92	5,21
		S3p	>30		
WK	A	S1p	6,65		
		S2p	5,71	5,01	7,61
		S3p	6,63		
	B	S1p	5,53	5,31	6,32
		S2p	5,75		
		S3p	6,09		
TUR	A	S1p	23,86		
		S2p	16,12	10,95	13,80
		S3p	16,66		
	B	S1p	26,96		
		S2p	27,63		
		S3p	26,43	10,92	10,50

Źródło: obliczenia własne.

W przypadku prognoz ekstrapolacyjnych modele hierarchiczne okazały się dla wszystkich zmiennych bardziej efektywne od modeli sieci neuronowych. Porównanie dokładności prognoz otrzymanych na podstawie sieci neuronowych oraz modeli oszczędnych dla zmiennych *M* i *TUR* wypada na korzyść modeli oszczędnych. Dla zmiennej *WK* bardziej efektywne były prognozy dla sieci neuronowych.

4. Podsumowanie

Z przeprowadzonych w pracy analiz porównawczych dokładności prognoz interpolacyjnych dla szeregów z lukami w danych wynika, że jedynie dla zmiennej *WK*, charakteryzującej się najmniejszym rozstępem wskaźników sezonowości, najniższe błędy prognoz otrzymano na podstawie sieci neuronowych. Okazały się one natomiast mniej efektywne w prognozowaniu zmiennych o większym zróżnicowaniu tych wskaźników. W prognozowaniu ekstrapolacyjnym najbardziej dokładne były prognozy otrzymane na podstawie modeli hierarchicznych. Zwiększeniu liczby luk w danych towarzyszył za-

zwyczaj wzrost błędów względnych obu rodzajów prognoz. Spostrzeżenie to odnosi się zwłaszcza do modeli sieci neuronowych.

Literatura

Dintcheva-Biś D., *Zastosowanie sieci neuronowych do prognozowania szeregów czasowych*, Prace Instytutu Ekonometrii i Statystyki UŁ, Seria C, Nr 124, Łódź 1999.

Ekonometryczne metody predykcji dla danych sezonowych w warunkach braku pełnej informacji, red. J. Zawadzki, US, Szczecin 1999.

Markiewicz A., *Zastosowanie modeli hierarchicznych oraz sieci neuronowych w prognozowaniu zjawisk ekonomicznych*, Akademia Rolnicza w Szczecinie, Szczecin 2004 (maszynopis rozprawy doktorskiej).

Zastosowanie hierarchicznych modeli szeregów czasowych w prognozowaniu zmiennych ekonomicznych z wahaniami sezonowymi, red. J. Zawadzki, AR, Szczecin 2003.

EMPIRICAL COMPARATIVE ANALYSIS OF ACCURACY OF SELECTED FORECASTING METHOD IN SITUATION OF LACK OF FULL INFORMATION

Summary

The article presents results of missing data forecasting in economical time series with seasonal fluctuations. Accuracy of inter- and extrapolation forecasts was compared by using predictors based on time series models with trigonometrical polynomial, hierarchical and neural networks. The variables with different intensity of seasonality were used as samples.