Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 295 · 2016

Monika Miśkiewicz-Nawrocka

Uniwersytet Ekonomiczny w Katowicach
Wydział Zarządzania
Katedra Matematyki
monika.miskiewicz@ue.katowice.pl

WPŁYW LICZBY „NAJBLIŻSZYCH SĄSIADÓW" NA DOKŁADNOŚĆ PROGNOZ EKONOMICZNYCH SZEREGÓW CZASOWYCH

Streszczenie: Metoda najbliższych sąsiadów jest jedną z metod prognozowania szeregów czasowych. W metodzie tej, prognozę (N+1)-go elementu $\hat{x}_{N+1}$ szacuje się jako średnią ważoną obserwacji x_{i+1}, gdzie wektory x_i^d są k najbliższymi sąsiadami wektora x_N^d w zrekonstruowanej d-wymiarowej przestrzeni stanów. Istotnym problemem podczas stosowania tej metody jest wyznaczenie prawidłowej liczby najbliższych sąsiadów, która powinna być brana pod uwagę przy wyznaczaniu prognoz. Głównym celem artykułu jest zbadanie wpływu liczby najbliższych sąsiadów na dokładność prognoz ekonomicznych szeregów czasowych. Badania zostały przeprowadzone w oparciu o wybrane finansowe szeregi czasowe.

Słowa kluczowe: metoda najbliższych sąsiadów, prognozowanie szeregów czasowych, rekonstrukcja przestrzeni stanów.

Wprowadzenie

W teorii nieliniowych układów dynamicznych do prognozowania przyszłych wartości szeregów czasowych można zastosować metodę analogową – zwaną metodą „najbliższych sąsiadów". W metodzie tej, prognozowaną wartość szeregu x_n ustala się na podstawie średniej ważonej pierwszych współrzędnych punktów, będących najbliższymi (w sensie odległości euklidesowej) sąsiadami punktu x_n^d w zrekonstruowanej przestrzeni stanów, odpowiadającemu obserwacji x_n. Istotnym problemem przy stosowaniu metody analogowej jest ustalenie prawidłowej liczby najbliższych sąsiadów – punktów zrekonstruowanej przestrzeni stanów, które należy wziąć pod uwagę przy wyznaczeniu prognozy.

W pracy został zbadany wpływ liczby najbliższych sąsiadów na dokładność otrzymanych prognoz zjawisk ekonomicznych, opisanych za pomocą ekonomicznych szeregów czasowych. Badania empiryczne zostały przeprowadzone na podstawie rzeczywistych, ekonomicznych szeregów czasowych. W celu przeprowadzenia obliczeń wykorzystano program napisany przez autorkę w języku programowania Delphi oraz arkusz kalkulacyjny Excel.

1. Rekonstrukcja przestrzeni stanów układu dynamicznego

Rekonstrukcja przestrzeni stanów polega na odtworzeniu, jedynie na podstawie jednowymiarowego szeregu obserwacji, przestrzeni stanów układu dynamicznego. Jedną z metod rekonstrukcji jest metoda opóźnień, wprowadzona niezależnie przez N.H. Packarda [Packard i in., 1980] oraz F. Takensa [1981]. Rekonstrukcja przestrzeni odbywa się poprzez zanurzenie szeregu czasowego w przestrzeni o wyższym wymiarze, tj. poprzez odtworzenie trajektorii układu w wielowymiarowej przestrzeni wektorowej. Elementami tej przestrzeni są d-wymiarowe wektory zwane d-historiami, które powstają w wyniku przesunięcia oryginalnego szeregu czasowego o pewną stałą wartość opóźnienia czasowego τ [Kantz, Schreiber, 2004]:

$$s_t^d = \left(s_t, s_{t-\tau}, \ldots, s_{t-(d-1)\tau}\right) \tag{1}$$

gdzie: $(d-1)\tau + 1 \le t \le N$

s_t – obserwacje oryginalnego szeregu,

d – wymiar rekonstruowanej przestrzeni (zwany również wymiarem zanurzenia),

τ – opóźnienie czasowe.

F. Takens udowodnił, że dla $d \ge 2m + 1$, gdzie m jest wymiarem atraktora, a d – wymiarem zanurzenia, zrekonstruowana przestrzeń stanów układu jest topologicznie równoważna z „oryginalną" przestrzenią układu dynamicznego [Zawadzki, 1996].

2. Metoda najbliższych sąsiadów

Metoda najbliższych sąsiadów – NS, zwana też metodą analogową, została zaproponowana przez E.N. Lorenza [1969] i jest najstarszą metodą prognozowania chaotycznych szeregów czasowych. Jej podstawą teoretyczną jest fakt, iż stany układów deterministycznych ewoluują w czasie w podobny sposób. W przypadku szeregów czasowych, gdy nie znamy funkcji f, opisującej dynami-

kę układu i dysponujemy jedynie jednowymiarowym szeregiem obserwacji, należy przeprowadzić rekonstrukcję przestrzeni stanów według wzoru (1). Jeśli $s_{t_0}^d$ jest najbliższym sąsiadem punktu s_N^d, to również $f_T\left(s_N^d\right) \approx f_T\left(s_{t_0}^d\right)$, a stąd wynika, że $s_{N+T} \approx s_{t_0+T}$. Zatem wartość s_{t_0+T} można przyjąć jako prognozę obserwacji s_{N+T} analizowanego szeregu czasowego [Lorenz, 1969, s. 51; Nowiński, 2007].

W metodzie najbliższych sąsiadów prognozę dla $N + 1$ elementu $\hat{s}_{N+1}$ szacuje się jako średnią ważoną obserwacji s_{i+1}, gdzie wektory s_i^d są k najbliższymi sąsiadami wektora s_N^d w zrekonstruowanej d-wymiarowej przestrzeni stanów:

$$\hat{s}_{N+1} = \sum_{i=1}^{k} w_i s_{i+1} \tag{2}$$

gdzie: s_{i+1} – pierwsza współrzędna wektora s_{i+1}^d

$w_i = w\left(\left\| s_N^d - s_i^d \right\|\right)$ – waga i-tego sąsiada wektora s_N^d

$w : R \to R$ jest dowolną funkcją malejącą spełniającą warunki:

$$w_i = w\left(\left\| s_N^d - s_i^d \right\|\right) > 0$$

$$\sum_{i=1}^{k} w_i = \sum_{i=1}^{k} w\left(\left\| s_N^d - s_i^d \right\|\right) = 1$$

$$i = 1, 2, \ldots, k$$

Wagi są dobierane w ten sposób, by bliżsi sąsiedzi mieli większy wpływ na otrzymaną prognozę. Stąd wagę i-tego sąsiada można wyznaczyć według wzorów [Orzeszko, 2005]:

$$w_i = \frac{1}{k-1}\left(1 - \frac{d_i}{\sum_{i=1}^{k} d_i}\right) \tag{3}$$

$$w_i = \frac{2(k+1-i)}{k(k+1)} \tag{4}$$

$$w_i = \frac{e^{-d_i}}{\sum_{i=1}^{k} e^{-d_i}} \tag{5}$$

gdzie: $d_i = \left\| s_N^d - s_i^d \right\|$ oznacza odległość między wektorami s_N^d i s_i^d, $i = 1, 2, \ldots, k$.

3. Badania empiryczne

Przedmiotem badania były logarytmy dziennych stóp zwrotu indeksów giełd światowych: NIKKEI225 – indeks na Giełdzie Papierów Wartościowych w Tokio (NKX), S&P500 – indeks giełdy w Nowym Jorku (SPX) i WIG – indeks na Giełdzie Papierów Wartościowych w Warszawie; kursów walut: euro (EUR) i jena japońskiego (JPY) wobec złotego; cen akcji spółek: ING Bank Śląskiego (BSK) i Żywca (ZWC) oraz cen surowców: ropy naftowej (SC), srebra (XAG) i złota (XAU), postaci:

$$x_t = \ln s_t - \ln s_{t-1} \tag{6}$$

gdzie: s_t – obserwacja szeregu, notowane w okresie 3.01.2000r.-26.08.2013r[1].

W pierwszym etapie badania, dla wybranych szeregów czasowych oszacowano parametry rekonstrukcji przestrzeni stanów metodą opóźnień: stosując funkcję autokorelacji – *ACF* [Ramsey i in., 1990], oszacowano czas opóźnień τ, natomiast za pomocą metody najbliższego pozornego sąsiada – *FNN* [Abarbanel i in., 1992], obliczono wymiar zanurzenia d (tab. 1).

Tabela 1. Parametry rekonstrukcji przestrzeni stanów analizowanych szeregów czasowych

Szereg	τ	d	Szereg	τ	d
EUR	21	8	SPX	15	7
ING	6	8	WIG	16	7
JPY	2	6	XAG	4	8
NKX	6	6	XAU	22	7
SC	2	6	ZWC	17	10

W celu zbadania wpływu liczby najbliższych sąsiadów na dokładność prognoz oszacowanych metodą najbliższych sąsiadów, wyznaczono prognozy analizowanych szeregów dla horyzontu prognozy T = 1, 2, …, 10. W procesie prognozowania jako liczbę najbliższych sąsiadów przyjęto k = 2, 3, …, 100. Jako wagi i-tego sąsiada przyjęto: średnią arytmetyczną pierwszych współrzędnych najbliższych sąsiadów [Abarbanel i in., 1992] (NS_A), wagi zadane wzorem (3) – NS_B, wzorem (4) – NS_C i wzorem (5) – NS_D.

[1] Dane pochodzą z archiwum plików strony internetowej stooq.com [www 1].

Oceny trafności wyznaczonych prognoz dokonano za pomocą pierwiastka błędu średniokwadratowego (*RMSE*):

$$\sigma_T = \sqrt{\frac{1}{h}\sum_{t=n+1}^{n+h}\left(s_T - \hat{s}_T\right)^2} \tag{7}$$

gdzie: s_T – rzeczywista wartość badanej zmiennej w momencie T,
$\hat{s}_T$ – prognoza wartości zmiennej w momencie T, $T = n + 1, \ldots, n + h$,
h – liczba naturalna, oznaczająca odległość okresu prognozowanego od okresu bieżącego.

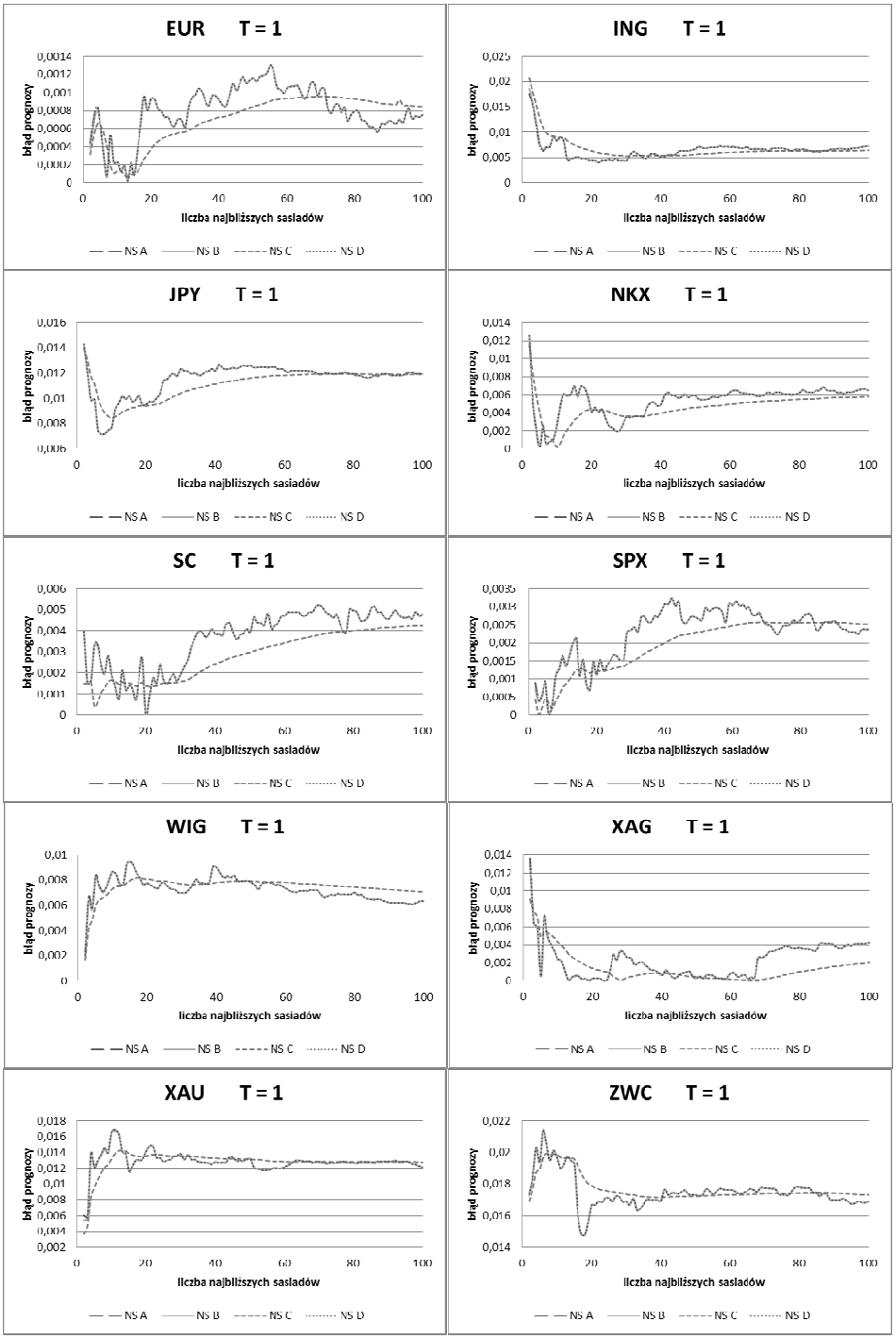
Na rysunku 1 przedstawiono błędy uzyskanych prognoz dla horyzontu prognozy T = 1, w zależności od liczby najbliższych sąsiadów, zastosowanych w procesie prognozowania metoda najbliższych sąsiadów, natomiast rys. 2 prezentuje zależność pomiędzy błędami predykcji, w całym przedziale weryfikacji, dla horyzontu prognozy T = 10 a liczbą najbliższych sąsiadów.

Analizując otrzymane wyniki, można stwierdzić, iż błędy prognoz otrzymane metodami NS_A, NS_B i NS_D przyjmują zbliżone wartości. Sytuacja ta została przedstawiona na rys. 1 oraz 2, na których wykresy zależności pomiędzy błędem prognozy a liczbą najbliższych sąsiadów prawie pokrywają się dla wspomnianych metod. Natomiast błędy prognoz, uzyskane metodą NS_C wyraźnie różnią się od pozostałych.

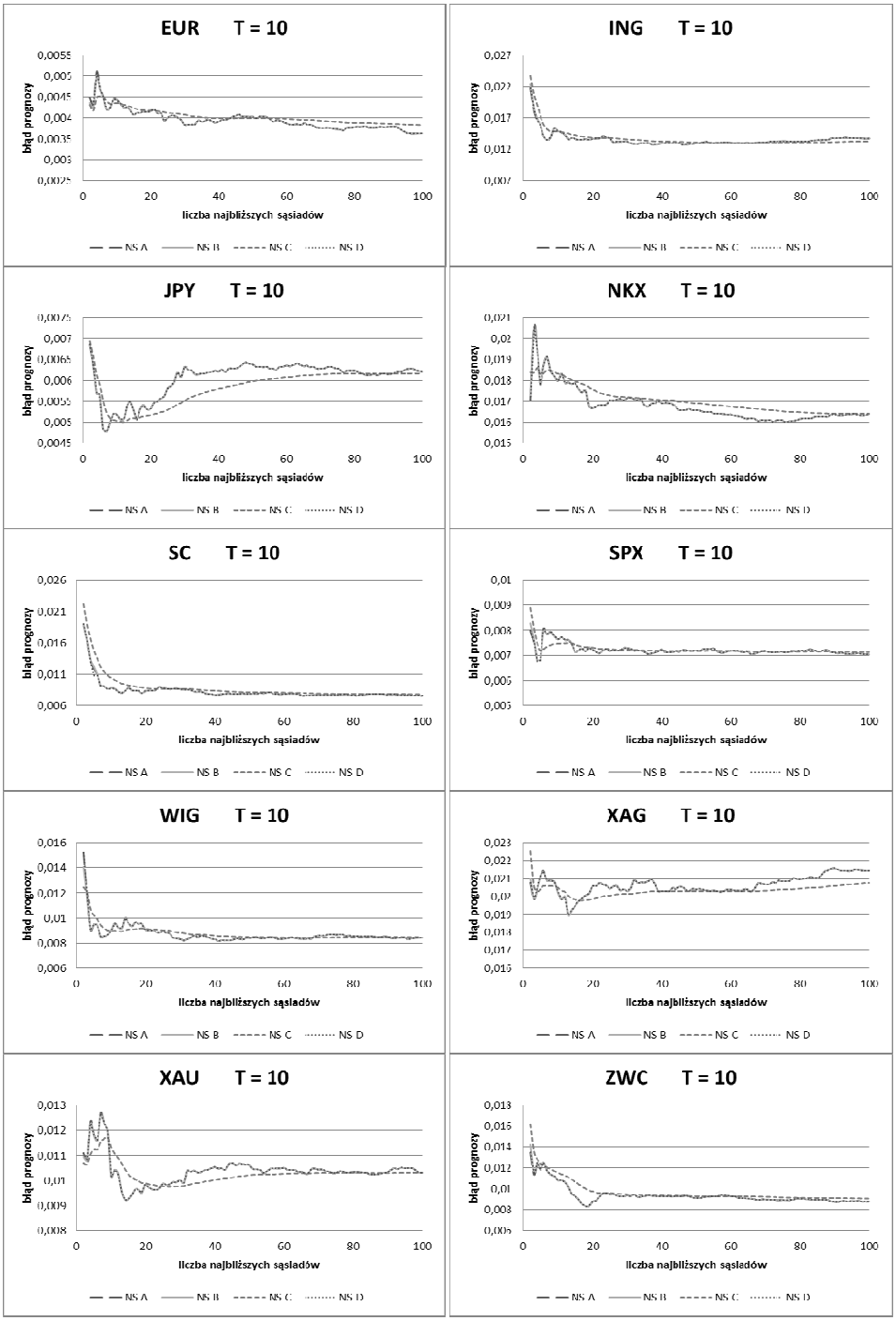
Na podstawie przeprowadzonych badań wynika, że istnieje wyraźna zależność pomiędzy zastosowaną w procesie prognozowania liczbą najbliższych sąsiadów a dokładnością otrzymanych prognoz, zarówno dla horyzontu prognozy T = 1, jak i w całym przedziale weryfikacji dla T = 10.

Tabela 2. Optymalna liczba najbliższych sąsiadów dla horyzontu prognozy T = 1

Szereg	***kmin* dla T = 1**			
	NS_A	NS_B	NS_C	NS_D
EUR	13	13	13	13
ING	22	22	30	22
JPY	7	7	10	7
NKX	5	5	10	5
SC	20	20	5	20
SPX	6	6	3	6
WIG	2	2	2	2
XAG	19	19	67	19
XAU	3	3	2	3
ZWC	18	18	2	18



Rys. 1. Wpływ liczby najbliższych sąsiadów na błąd prognozy dla horyzontu prognozy T = 1



Rys. 2. Wpływ liczby najbliższych sąsiadów na błąd prognozy w całym przedziale weryfikacji

Dla horyzontu prognozy T = 1 najmniejsze błędy prognozy otrzymano dla liczby najbliższych sąsiadów $k \leq 20$. Wyjątek stanowi szereg ING oraz szereg XAG dla metody NS_C. W tabeli 2 przedstawiono ilość najbliższych sąsiadów, dla których otrzymano najmniejsze błędy prognoz dla horyzontu prognozy T = 1. Wraz ze wzrostem liczby najbliższych sąsiadów ($k > kmin$) wartość błędu prognozy rośnie, a następnie stabilizuje się na pewnym poziomie (ING, JPY, SPX, XAU, ZWC). Najdokładniej widać to dla metody NS_C, gdzie wagi najbliższych sąsiadów były ustalane według wzoru (4), w którym pod uwagę bierze się numer *i*-tego sąsiada.

Badając dokładność wyznaczonych prognoz w całym przedziale weryfikacji dla horyzontu prognozy T = 10, można stwierdzić, iż dla większości badanych szeregów (EUR, ING, NKX, SC, SPX, WIG, ZWC) wartości błędu prognozy maleją wraz ze wzrostem liczby *k* najbliższych sąsiadów i stabilizują się (ING, SC, SPX, WIG) na pewnym poziomie dla $k > 40$. Dla szeregów ING, SC, SPX i WIG błędy prognoz stabilizują się na poziomie odpowiednio 0,013, 0,0076, 0,007 i 0,0084. Dla analizowanych szeregów liczbę najbliższych sąsiadów, dla których uzyskano najmniejsze błędy prognoz zamieszczono w tab. 3.

Tabela 3. Optymalna liczba najbliższych sąsiadów w całym przedziale weryfikacji dla horyzontu prognozy T = 10

Szereg	***kmin* dla T = 1**			
	NS_A	NS_B	NS_C	NS_D
EUR	100	100	100	100
ING	46	46	67	46
JPY	7	7	13	7
NKX	76	76	100	76
SC	100	100	100	100
SPX	4	4	100	4
WIG	41	41	68	41
XAG	13	13	19	13
XAU	14	14	27	14
ZWC	18	18	100	18

Analizując dane zawarte w tab. 3, można stwierdzić, że dla większości badanych szeregów dokładność prognoz rośnie wraz za wzrostem liczby najbliższych sąsiadów.

Podsumowanie

W opracowaniu zbadano wpływ liczby najbliższych sąsiadów, zastosowanych w procesie prognozowania metodą najbliższych sąsiadów na dokładność uzyskanych prognoz. Badania empiryczne przeprowadzono na przykładzie szeregów logarytmów dziennych stóp zwrotu notowań NIKKEI225, S&P500, WIG, euro, jena japońskiego, ING Banku Śląskiego, Żywca oraz cen ropy naftowej, srebra i złota. Przeprowadzone badania pokazują, że przyjęta liczba najbliższych sąsiadów, stosowanych w rozważanej metodzie prognozowania w bardzo istotny sposób wpływa na dokładność otrzymanych prognoz. Analizując otrzymane wyniki, można stwierdzić, że dla horyzontu prognozy T = 1 najmniejsze błędy prognoz uzyskano dla liczby najbliższych sąsiadów nie większej niż 20. Wyjątek stanowił szereg ING oraz XAG (metoda NS_C). Natomiast w całym przedziale weryfikacji dla horyzontu prognozy T = 10, w większości badanych szeregów wraz ze wzrostem liczby najbliższych sąsiadów błędy prognoz maleją lub zaczynają się stabilizować na pewnym poziomie. Pozwala to wnioskować, że począwszy od pewnej wartości liczby najbliższych sąsiadów *k*, różnice pomiędzy błędami prognoz są bardzo niewielkie (coraz mniej istotne), a więc zwiększanie liczby *k* w zasadzie nie prowadzi już do poprawy dokładności otrzymanych prognoz.

Literatura

Abarbanel H.D., Brown R., Kennel M.B. (1992), *Determining Embedding Dimension for Phase Space Reconstruction Using a Geometrical Construction*, "Physical Review A", Vol. 45(6).

Kantz H., Schreiber T. (2004), *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge.

Lorenz E.N. (1969), *Atmospheric Predictability as Revealed by Naturally Occurring Analogues*, "J. Atmos. Sci.", Vol. 26.

Nowiński M. (2007), *Nieliniowa dynamika szeregów czasowych*, Wydawnictwo Akademii Ekonomicznej, Wrocław.

Orzeszko W. (2005), *Identyfikacja i prognozowanie chaosu deterministycznego w ekonomicznych szeregach czasowych*, Polskie Towarzystwo Ekonomiczne, Warszawa.

Packard N.H., Crutchfield J.P., Farmer J.D., Shaw R.S. (1980), *Geometry from a Time Series*, "Physical Review Letters", Vol. 45.

Ramsey J.B., Sayers C.L., Rothman P. (1990), *The Statistical Properties of Dimension Calculations Using Small Data Sets: Some Economic Applications*, "International Economic Review", Vol. 31, No. 4.

Takens F. (1981), *Detecting Strange Attractors in Turbulence* [w:] D.A. Rand, L.S. Young (ed.), *Lecture Notes in Mathematics*, Springer, Berlin.

Zawadzki H. (1996), *Chaotyczne systemy dynamiczne*, Wydawnictwo Akademii Ekonomicznej w Katowicach, Katowice.

[www 1] stooq.com (dostęp: 1.09.2013).

EFFECT OF THE NUMBER OF "NEAREST NEIGHBORS" ON THE ACCURACY OF ECONOMIC TIME SERIES FORECASTS

Summary: One of time series forecasting method is the nearest neighbors method. In this method, the forecast for (N+1)-th element $\hat{x}_{N+1}$ is estimated as a weighted average of observations x_{i+1}, where the vectors x_i^d are k nearest neighbors of vector x_N^d in the reconstructed d-dimensional state space. An important problem when using nearest neighbors method is to determine the correct number of nearest neighbors, that should be taken into account in the determination of forecasts. The aim of the article will be to research the effect of the number of nearest neighbors on the accuracy of economic time series forecasts. The test will be conducted on the basis of selected financial time series.

Keywords: the nearest neighbors method, time series forecasting, state space reconstruction.