

Jacek Szandula

Akademia Ekonomiczna we Wrocławiu

PRÓBA OKREŚLENIA SPOSOBU OCENY DOPUSZCZALNOŚCI PROGNOZ WYZNACZANYCH METODĄ NAJMNIEJSZEGO SYMPLEKSU

1. Wstęp

Do istotnych cech procesu prognostycznego należy zaliczyć to, że wyznaczenie prognozy z reguły nie kończy prac badawczych. Odbiorcy prognozy zależy bowiem nie tylko na jej znajomości, lecz ze względów praktycznych istotna jest również informacja o stopniu zaufania do prognozy. Prognozy, których stopień niepewności jest akceptowany przez odbiorcę nazywa się dopuszczalnymi [Dittmann 2004, s. 32]. Ocenę dopuszczalności prognoz, o ile to możliwe, przeprowadza się na podstawie analizy błędów predykcji. W przypadku prognozowania ekonometrycznego zagadnienie to jest w pełni wyjaśnione (zob. np. [Theil 1979]). Określanie dopuszczalności prognoz wyznaczanych metodą najmniejszego sympleksu przysparza jednak pewnych trudności.

Celem artykułu jest przedstawienie propozycji sposobu oceny dopuszczalności prognoz wyznaczanych metodą najmniejszego sympleksu.

W pierwszej części artykułu opisana została idea metody. Następnie poruszono problematykę oceny dopuszczalności prognoz. Zaproponowane zostały sposoby oceny dopuszczalności prognoz. W dalszej części przedstawiono eksperyment, którego celem była próba weryfikacji wcześniej rozważanych sposobów oceny dopuszczalności prognoz.

2. Idea metody

Metoda zaproponowana przez G. Sugiharę i R.H. Maya [1990] została zaprojektowana w celu prognozowania szeregów czasowych charakteryzujących się dużymi wahaniem przypadkowymi. Jednak jak wykazał W. Drosdowsky [1994], metoda ta jest także użyteczna w prognozowaniu szeregów czasowych, w których w pew-

nych odstępach czasowych, zwykle nieregularnych, pojawiają się określone prawidłowości. Oznacza to, że założenie o znajomości mechanizmu generującego szereg czasowy, które często jest stosowane w przypadku prognozowania szeregów czasowych, ustępuje tu założeniu o podobieństwie rozwoju. Z tego względu metodę tę można zaliczyć do metod analogowych (zob. np. [Cieślak 2005]).

Istotą metody jest przedstawienie kolejnych k -elementowych fragmentów szeregu czasowego jako punktów w przestrzeni k -wymiarowej. K -elementowy fragment szeregu czasowego, który zostanie tu określony środkiem ciężkości sympleksu, w przestrzeni k -wymiarowej ma współrzędne:

$$y_n = [y_{n-(k-1)g} \ \dots \ y_{n-2g} \ y_{n-g} \ y_n]^T, \quad (1)$$

gdzie: y_n – środek ciężkości sympleksu,

g – wielkość odstępów pomiędzy kolejnymi obserwacjami – zwykle $g = 1$.

W analogiczny sposób jako k -wymiarowe punkty przedstawiane są pozostałe fragmenty szeregu czasowego:

$$x_i = [x_{n-i-(k-1)g} \ \dots \ x_{n-i-2g} \ x_{n-i-g} \ x_{n-i}]^T, \quad (2)$$

gdzie: x_i – punkt reprezentujący i -ty (licząc w kierunku od obserwacji najmłodszych do najstarszych) fragment szeregu czasowego.

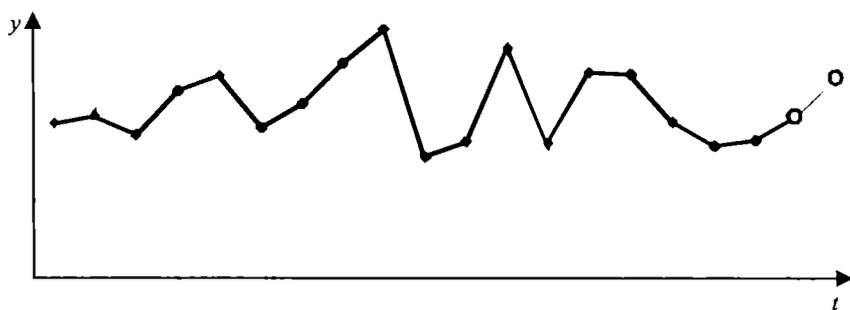
$i = 1, 2, \dots, i \leq n-k$.

Pośród „sąsiadów” punktu y_n wybierana jest kombinacja $k+1$ punktów tworzących sympleks o najmniejszej średnicy zawierający punkt y_n . Jeżeli prognozowany punkt y_n znajduje się na wierzchołku, na krawędzi lub ścianie sympleksu, wówczas do prognozowania można wybrać sympleks o mniejszej niż k liczbie wymiarów, tzn. zbiór punktów $\{x_i, \dots, x_j\}$ ulega redukcji.

Obliczenie prognozy wymaga wyznaczenia współrzędnych wierzchołków sympleksu po $T-n$ momentach/okresach (przez T rozumiany jest okres prognozowania). Na ich podstawie oraz przy użyciu współrzędnych barycentrycznych punktu y_n w znalezionym sympleksie ustalana jest wartość prognozy.

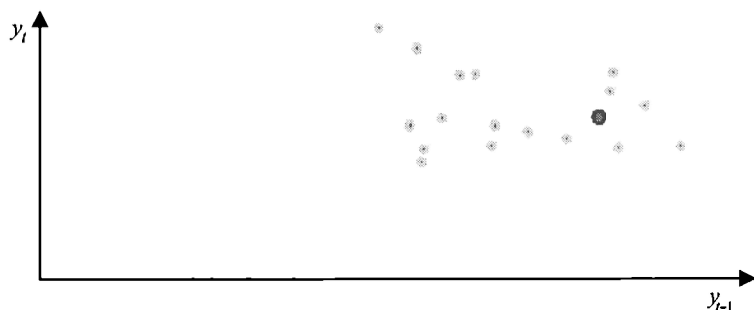
Poszczególne kroki metody najłatwiej prześledzić jest dla przypadku, gdy $k = 2$ i $g = 1$. Rysunek 1 prezentuje przykładowy szereg czasowy. Dla ustalonego parametru $k = 2$ wybiera się dwie ostatnie obserwacje. Wybrane wartości przedstawia się w przestrzeni dwuwymiarowej jako środek ciężkości y_n o współrzędnych $[y_{n-1}, y_n]$. Tak samo postępuje się z pozostałymi parami następujących po sobie obserwacji. W ten sposób otrzymuje się zbiór punktów otaczających środek ciężkości (rys. 2).

Z otrzymanego zbioru wybierane są trzy punkty (w ogólnym przypadku $k+1$ punktów) tworzące wierzchołki sympleksu (w tym przypadku jest nim trójkąt) o najmniejszej średnicy spośród sympleksów zawierających środek ciężkości. Przez średnicę rozumiana jest tu maksymalna odległość między punktami tworzącymi sympleks.



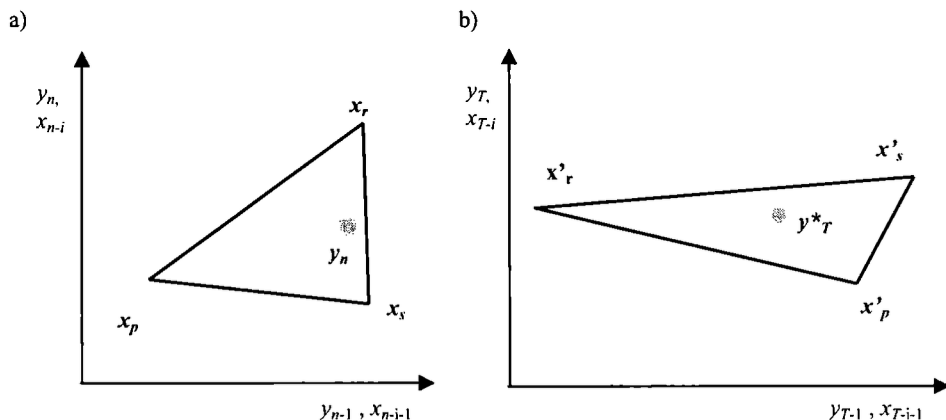
Rys. 1. Przykładowy szereg czasowy

Źródło: opracowanie własne.



Rys. 2. Szereg czasowy jako zbiór punktów w przestrzeni dwuwymiarowej

Źródło: opracowanie własne.

Rys. 3. Prognozowanie przy użyciu metody najmniejszego sympleksu w przypadku $k = 2$

Źródło: opracowanie własne.

Rysunek 3a przedstawia szukany sympleks $x_p x_r x_s$. Odległości punktu y_n od poszczególnych wierzchołków trójkąta określają współrzędne barycentryczne punktu y_n w sympleksie $x_p x_r x_s$. Aby wyznaczyć prognozę y^*_T , należy znaleźć wierzchołki nowego sympleksu $x'_p x'_r x'_s$, przy czym:

$$x'_i = x_{i-(T-n)}. \quad (3)$$

Po znalezieniu wierzchołków nowego sympleksu wyznacza się nowy środek ciężkości y^*_T , który ma współrzędne barycentryczne w sympleksie x'_p, x'_r, x'_s identyczne ze współrzędnymi barycentrycznymi punktu y_n w sympleksie x_p, x_r, x_s (rys. 3b). Ostatnia współrzędna punktu y^*_T wyznacza wartość prognozy na okres T .

3. Eksperyment

Model najmniejszego sympleksu można zapisać w postaci:

$$y = Xw, \quad (4)$$

gdzie: y – środek ciężkości,

X – macierz wierzchołków sympleksu,

w – wektor współrzędnych barycentrycznych.

W modelu tym nie ma mowy o błędzie oszacowania czy losowości. Nieznane z góry wartości współrzędnych barycentrycznych w są jednoznacznie wyznaczone na podstawie wzajemnego położenia środka ciężkości sympleksu y oraz wierzchołków sympleksu X (zob. np. [Borsuk 1977, s. 120]). Można stwierdzić, że model idealnie dopasowuje się do danych empirycznych. Nie można jednak zakładać, że sytuacja ta będzie się utrzymywać przy wyznaczaniu prognoz. Błędy prognoz z pewnością będą się pojawiać. Pytanie, w jaki sposób określać ich wielkość *ex ante*? Ponieważ nie jest możliwe bezpośrednie obliczenie wariancji błędów predykcji, zasadne jest rozpatrywanie alternatywnych sposobów oceny dopuszczalności. Autor proponuje wziąć pod uwagę:

- k -wymiarową miarę sympleksu,
- odległość środka ciężkości od wierzchołków sympleksu,
- ważone kwadraty różnic współrzędnych wierzchołków i środka ciężkości,
- wariancję współrzędnych wierzchołków sympleksu generujących prognozę,
- zmienną syntetyczną.

Pierwsze trzy sposoby związane są ze wzajemnym położeniem sympleksu oraz punktu ciężkości. Można powiedzieć, że wykorzystują one przeszłe własności używanego modelu. Czwarty sposób w pewnym sensie wybiega w okres prognozy, a wykorzystanie zmiennej syntetycznej stanowi próbę połączenia korzystnych efektów (polegających na możliwości oceny dopuszczalności prognozy) otrzymywanych wcześniej rozpatrywanymi sposobami.

k -wymiarową miarę sympleksu stanowi wartość bezwzględna następującego wyrażenia:

$$m = \frac{\pm 1}{k!} \begin{vmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{k+1,1} & x_{k+1,2} & \dots & x_{k+1,k} \end{vmatrix}. \quad (5)$$

2-wymiarowa miara sympleksu jest polem trójkąta, 3-wymiarowa miara sympleksu jest objętością czworościanu. Hipoteza badawcza brzmi: im większa k -wymiarowa miara sympleksu, tym większy błąd prognozy.

Sumę odległości środka ciężkości od wierzchołków sympleksu wyznacza się w następujący sposób:

$$d = \sum_{i=1}^{k+1} \sqrt{\sum_{t=1}^k (x_{i,t} - y_t)^2}. \quad (6)$$

W tym przypadku można sformułować następującą hipotezę: im większa odległość środka ciężkości od wierzchołków sympleksu, tym większych błędów prognoz należy się spodziewać.

Kolejny z rozpatrywanych sposobów oceny dopuszczalności prognoz polegał na wyznaczeniu wartości:

$$w = \sum_{t=1}^k \sum_{i=1}^{k+1} w_i (x_{i,t} - y_t)^2, \quad (7)$$

gdzie w – suma ważonych kwadratów odchyleń wierzchołków od środka ciężkości sympleksu.

W tym przypadku przyjęto hipotezę, że błąd prognoz rośnie w miarę wzrostu wyznaczonej miary.

Z kolei wariancję współrzędnych tworzących prognozę wyznacza się jak poniżej:

$$v = \sqrt{\frac{1}{k+1} \sum_{i=1}^{k+1} (x_{i,t+1} - y_{t+1}^*)^2}. \quad (8)$$

Hipoteza: im większa wariancja współrzędnych tworzących prognozę, tym większy błąd prognozy.

Jako ostatni ze sposobów oceny dopuszczalności prognoz rozpatrzono zmienną syntetyczną powstałą przez zsumowanie zestandaryzowanych wcześniej rozpatrywanych zmiennych. Ponieważ wszystkie poprzednio rozpatrywane zmienne były destymulantami, zatem w przypadku zmiennej syntetycznej również przyjęto wstępną hipotezę, że w miarę wzrostu wartości zmiennej syntetycznej rosnać będzie błąd prognozy.

W celu weryfikacji postawionych wyżej hipotez wygenerowano sto 3000-elementowych szeregów. Szeregi te były generowane według następującego wzoru:

$$y_{t,i} = a_i \sin \left(\frac{2\pi(t - \sum_{j=0}^{i-1} b_j)}{b_i} \right) + e_t, \quad (9)$$

gdzie: $y_{t,i}$ – obserwacja szeregu w t -tym okresie i i -tym cyklu,
 t – numer okresu, $t = 1, 2, 3, \dots, 3000$,
 i – numer cyklu, $i = 1, 2, \dots; i \leq 3000$,
 a_i – parametr związany z amplitudą cyklu generowany z rozkładu normalnego $N(0, 3)$,
 b_i – parametr związany z długością cyklu generowany z rozkładu jednostajnego dyskretnego z przedziału $[1, 20]$,
 e_t – szum generowany z rozkładu normalnego $N(0, 0,5)$.

Wygenerowane w ten sposób szeregi charakteryzują się cykliczną oscylacją wokół zera. Cykle są nieregularne. Nieregularność pojawia się zarówno w długościach cykli, jak i w ich amplitudach. W pracy Szandudy [2006] opisano przydatność metody najmniejszego sympleksu do prognozowania tego rodzaju szeregów.

Na podstawie pierwszego tysiąca obserwacji każdego z szeregów stworzono bibliotekę wzorów, tj. punkty w przestrzeniach k -wymiarowych. Każdy z otrzymanych w ten sposób punktów potencjalnie mógł być wierzchołkiem sympleksu zawierającego pewien środek ciężkości. Kolejne 2000 obserwacji posłużyło do określenia środków ciężkości sympleksów oraz do wyznaczenia błędów prognoz *ex post*.

Po wyznaczeniu prognoz oraz ich błędów *ex post* obliczono wartości miar rozpatrywanych do oceny dopuszczalności prognoz. Następnie tworzone dwukolumnowe macierze (zob. rys. 4). Pierwsza kolumna macierzy zawierała posortowane niemalejąco wyniki rozważanej miary uzyskane dla poszczególnych sympleksów, a druga odpowiadające im błędy prognoz *ex post*. W dalszej kolejności grupowano obserwacje obu kolumn macierzy po pięćdziesiąt obserwacji. Jeżeli liczba wyznaczonych prognoz nie dzieliła się bez reszty przez 50, to liczba przypadków w ostatniej grupie była równa:

- 50 plus wartość tej reszty, gdy reszta była mniejsza od 25, lub
- wartości tej reszty, gdy reszta była nie mniejsza od 25.

Dla tak uzyskanych grup wartości obliczano (zob. rys. 3):

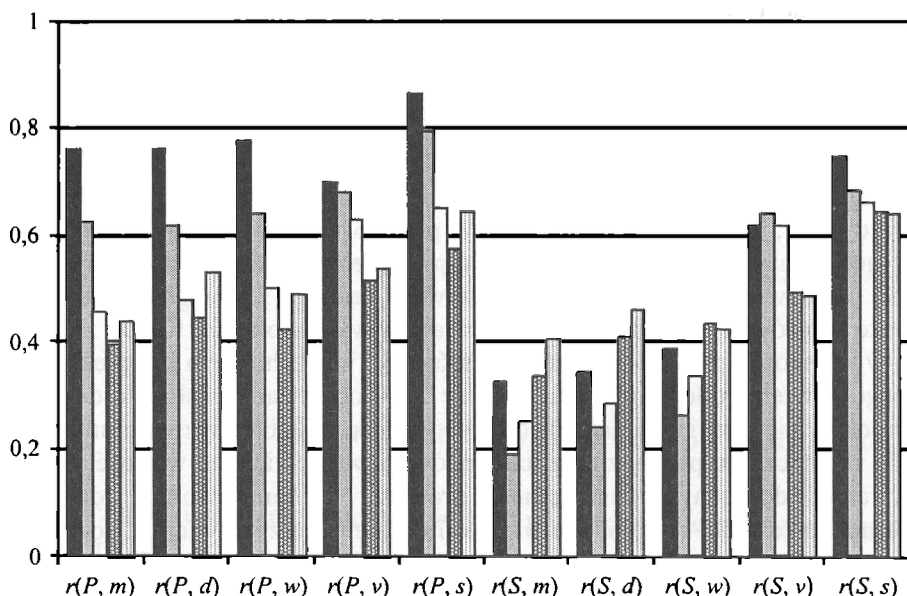
- dla kolumny pierwszej – środek przedziału klasowego,
- dla kolumny drugiej – odchylenie standardowe błędów prognoz *ex post*.

$$\begin{array}{l}
 \text{Środek przedziału klasowego } \bar{m}_{1-50} \left\{ \begin{array}{c} m_1 \quad q_1 \\ m_2 \quad q_2 \\ \vdots \\ m_{50} \quad q_{50} \\ m_{51} \quad q_{51} \\ \vdots \\ m_{100} \quad q_{100} \\ \vdots \end{array} \right\} \begin{array}{l} \text{odchylenie standardowe } \{q_1, \dots, q_{50}\} \\ \text{odchylenie standardowe } \{q_{51}, \dots, q_{100}\} \end{array} \\
 \text{Środek przedziału klasowego } \bar{m}_{51-100} \left\{ \begin{array}{c} m_1 \quad q_1 \\ m_2 \quad q_2 \\ \vdots \\ m_{50} \quad q_{50} \\ m_{51} \quad q_{51} \\ \vdots \\ m_{100} \quad q_{100} \\ \vdots \end{array} \right\}
 \end{array}$$

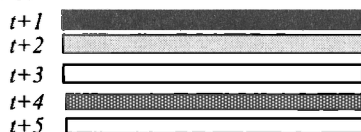
Rys. 4. Wyznaczanie wartości, na podstawie których obliczano korelacje

Źródło: opracowanie własne.

Na podstawie środków przedziałów klasowych oraz odchyłeń standardowych obliczono współczynniki korelacji Pearsona oraz korelacji rang Spearmana. Wartości średnie współczynników korelacji uzyskane na podstawie 100 szeregów dla parametru $k = 2, 3, 4, 5$ przedstawiają rys. 5-8.



Okres prognozy ($t =$ okres bieżący):

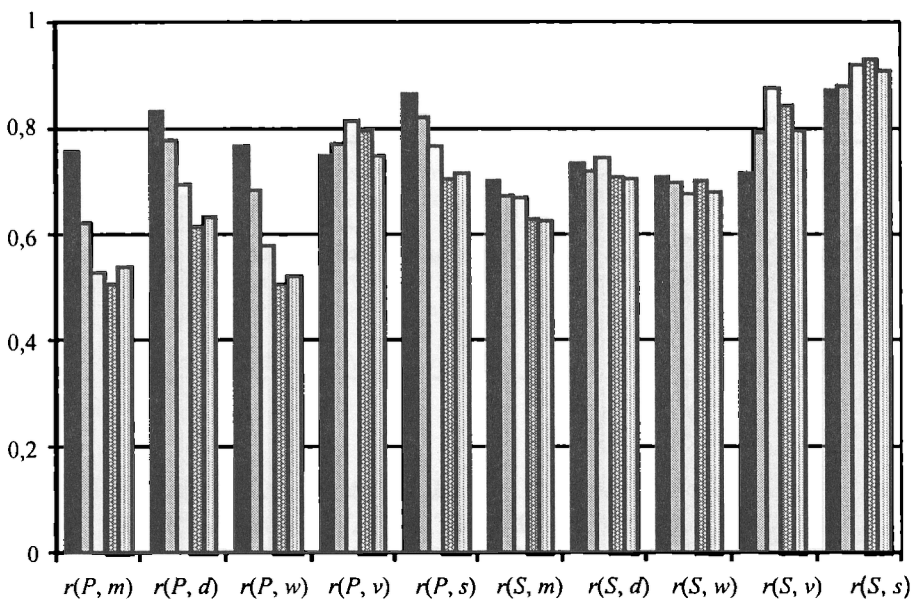


Legenda: Korelacje z odchyleniami standardowymi błędów prognoz *ex post*:

- $r(P, m)$ – współczynnik korelacji Pearsona dla miary sympleksu,
- $r(P, d)$ – współczynnik korelacji Pearsona dla odległości środka ciężkości od wierzchołków sympleksu,
- $r(P, w)$ – współczynnik korelacji Pearsona dla ważonych kwadratów różnic współrzędnych wierzchołków i środka ciężkości,
- $r(P, v)$ – współczynnik korelacji Pearsona dla wariancji współrzędnych wierzchołków sympleksu generujących prognozę,
- $r(P, s)$ – współczynnik korelacji Pearsona dla zmiennej syntetycznej,
- $r(S, m)$ – współczynnik korelacji Spearmana dla miary sympleksu,
- $r(S, d)$ – współczynnik korelacji Spearmana dla odległości środka ciężkości od wierzchołków sympleksu,
- $r(S, w)$ – współczynnik korelacji Spearmana dla ważonych kwadratów różnic współrzędnych wierzchołków i środka ciężkości,
- $r(S, v)$ – współczynnik korelacji Spearmana dla wariancji współrzędnych wierzchołków sympleksu generujących prognozę,
- $r(S, s)$ – współczynnik korelacji Spearmana dla zmiennej syntetycznej.

Rys. 5. Wartości średnie współczynników korelacji; $k = 2$

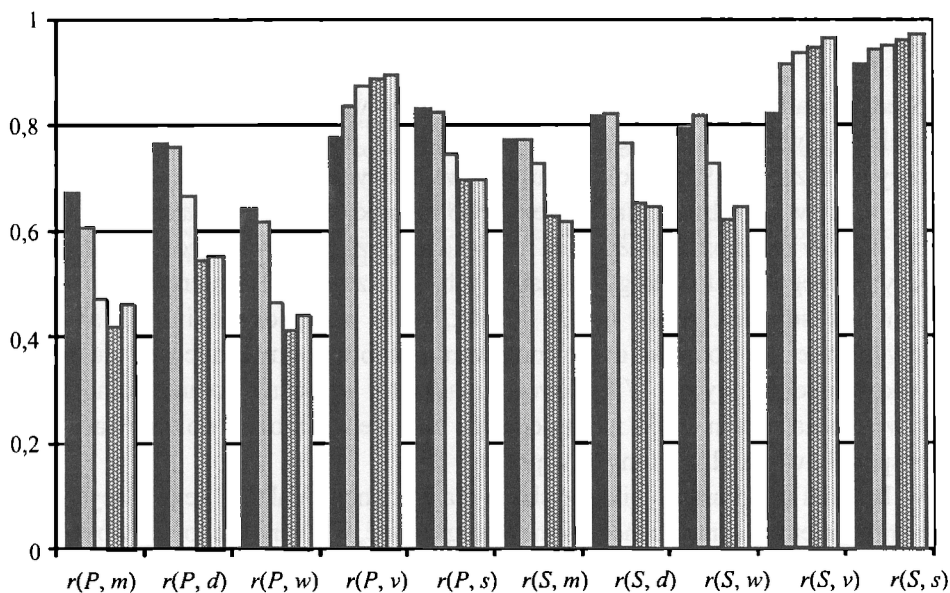
Źródło: opracowanie własne.



Legenda jak do rys. 5.

Rys. 6. Wartości średnie współczynników korelacji; $k = 3$

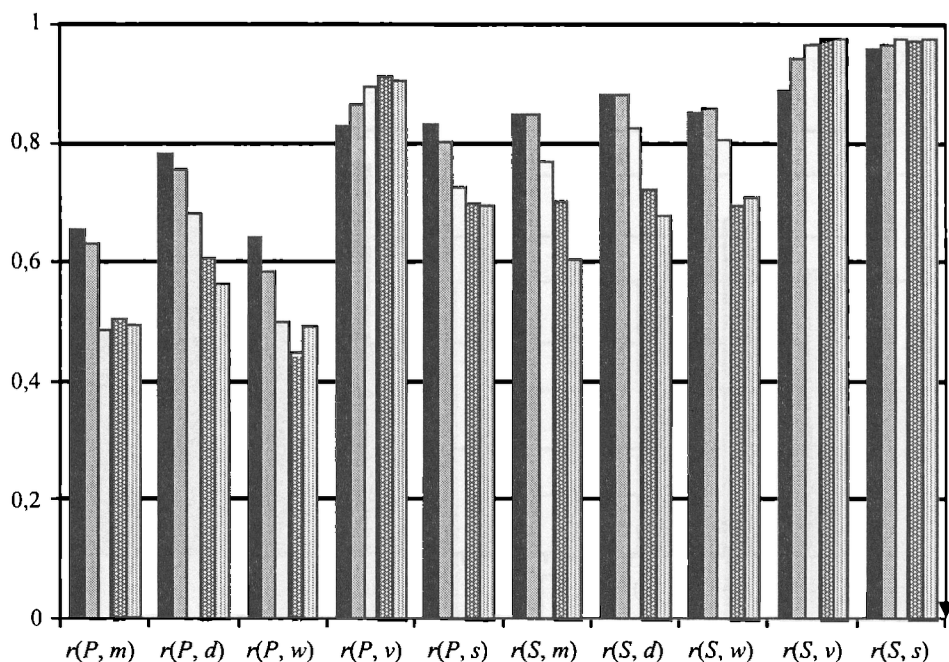
Źródło: opracowanie własne.



Legenda jak do rys. 5.

Rys. 7. Wartości średnie współczynników korelacji; $k = 4$

Źródło: opracowanie własne.



Legenda jak do rys. 5.

Rys. 8. Wartości średnie współczynników korelacji; $k = 5$

Źródło: opracowanie własne.

Średnie wartości współczynników korelacji we wszystkich analizowanych przypadkach były większe od zera. Na ogół wartości te przekraczały 0,6, niejednokrotnie były wyższe od 0,9. Wysokie wartości współczynników korelacji oznaczają, że nie ma podstaw do odrzucenia hipotez co do związków proponowanych miar z wariancją błędów prognoz. Należy mieć na uwadze to, że przeprowadzony eksperyment dotyczy jednego typu szeregów czasowych, tj. szeregów z nieregularnymi wahaniami okresowymi. Autor uważa jednak, że w przypadkach innych szeregów czasowych nie ma podstaw do formułowania hipotez odmiennych co do hipotez weryfikowanych powyżej. W praktyce oznacza to, że oceny dopuszczalności prognoz wyznaczanych metodą najmniejszego sympleksu można dokonywać za pomocą sposobów przedstawionych w pracy.

W badaniu najlepsze wyniki uzyskano dla wariancji punktów generujących prognozę oraz zmiennej syntetycznej, co może sugerować, iż miary te mogłyby być szczególnie przydatne do oceny dopuszczalności prognoz, aby jednak to potwierdzić, należałoby przeprowadzić eksperyment na większej próbie i w większych wymiarach.

Literatura

- Borsuk K., *Geometria analityczna wielowymiarowa*, PWN, Warszawa 1977.
- Cieślak M., *Prognozowanie analogowe*, [w:] *Prognozowanie gospodarcze. Metody i zastosowania*, red. M. Cieślak, Wydawnictwo Naukowe PWN, Warszawa 2005.
- Dittmann P., *Prognozowanie w przedsiębiorstwie*, Oficyna Ekonomiczna, Kraków 2004.
- Drozdowski W., *Analogue (non-linear) Forecasts of the Southern Oscillation Index Time-Series*, Weather Forecasting 1994, 78-84.
- Sugihara G., May R.M., *Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time Series*, Nature 1990, 344, 734-741.
- Szandula J., *Prognozowanie szeregów czasowych z nieregularnymi wahaniami okresowymi*, maszynopis, 2006.
- Theil H., *Zasady ekonometrii*, PWN, Warszawa 1979.

THE TRIAL OF ESTABLISHING THE WAY OF FORECASTS ACCEPTANCE EVALUATION FOR FORECASTS BUILT BY THE SMALLEST SIMPLEX METHOD

Summary

The paper considers the theme of time series forecasting using the smallest simplex method proposed in 1990 by G. Sugihara and R.M. May. The idea of the method is described. There is conducted a trial of establishing the way of forecasts acceptance evaluation in the paper. Because of the lack of checking possibilities of forecasts error variances, alternative approach is regarded. There is set an introductory hypothesis of possibilities benefit from others ways of forecasts acceptance evaluation. Among ways taken into account are:

- k -dimensional simplex measure,
- distances between simplex vertices and its gravity center,
- weighted differences between coefficients of simplex vertices and its gravity center,
- variance of coefficients generating forecast,
- theoretical variable consisted of above measures.

The author proposes and presents in practice the procedure of verifying possibilities of use of considered ways of forecasts acceptance evaluation.