

Jadwiga Suchecka

Uniwersytet Łódzki

Jan Kowalik

Politechnika Częstochowska

ANALIZA DANYCH CZASOWYCH Z OBSERWACJAMI NIETYPOWYMI Z WYKORZYSTANIEM METOD GEOSTATYSTYKI

1. Wstęp

Często wśród obserwacji ekonomicznych szeregów czasowych można wyróżnić wartości wyraźnie różniące się od pozostałych. Wartości takie określa się zwykle mianem obserwacji nietypowych (*outliers*). Przyczyny pojawienia się tego typu danych są różne, generalnie można ująć je w dwie grupy:

- błędne obserwacje, powstające najczęściej w procesie ich przetwarzania lub przenoszenia na różne nośniki danych,
- poprawne obserwacje, będące efektem nietypowych zdarzeń.

Obserwacje nietypowe z reguły powodują odkształcanie typowych tendencji i wymagają stosowania odpowiednich metod analizy jakościowej i ilościowej. Dlatego też analiza danych pod kątem występowania tego typu obserwacji oraz tworzenie procedur, które będą wykrywały i usuwały obserwacje nietypowe, jest zagadnieniem niezwykle ważnym.

Zasadniczym celem niniejszego opracowania jest przedstawienie teoretycznych aspektów wykorzystania geostatystycznej techniki krigingu w iteracyjnej procedurze wykrywania obserwacji nietypowych w szeregach czasowych, opartej na teście współczynnika prawdopodobieństwa. Rozważania ograniczono do modelu ARIMA (p, d, q).

Najogólniej definiując, geostatystyka jest działem statystyki, który dostarcza wielu skutecznych i powszechnie uznanych metod opisu ciągłości przestrzennej,

znajdujących początkowo zastosowanie w analizie danych przestrzennych, a obecnie również i czasowych.

Metoda krigingu jest podstawową, wykorzystywaną w geostatystyce, techniką estymacji, bazującą na analizie semiwariogramów. Kriging jest metodą dostarczającą najlepszych, liniowych, nieobciążonych estymatorów. Modele krigingu mogą być wykorzystane w statystyce w celu przewidywania wartości jednej bądź więcej realizacji zmiennych w danym punkcie w czasie i przestrzeni na podstawie danych wartości realizacji zmiennej w innych punktach.

2. Wykrywanie obserwacji nietypowych w szeregach czasowych

Wykrywanie obserwacji nietypowych w szeregu czasowym było po raz pierwszy badane przez Foxa [Fox 1972], który zdefiniował dwa rodzaje takich obserwacji zanieczyszczających szereg:

- addytywne obserwacje nietypowe (*additive outlier* – AO);
- innowacyjne obserwacje nietypowe (*innovative outlier* – IO).

Szereg czasowy, który nie zawiera żadnych obserwacji nietypowych, zdefiniowano jako szereg wolny od obserwacji nietypowych. Fox zaproponował również użycie testu współczynnika prawdopodobieństwa do wykrywania tych obserwacji. Propozycja Foxa dotyczyła modelu autoregresyjnego $AR(p)$ dla nieobserwowanego szeregu czasowego, wolnego od obserwacji nietypowych.

Dla zobrazowania istoty procedury przyjmujemy następujące założenia: zakładamy, że $\{Z_t\}$ będzie obserwowanym szeregiem czasowym zawierającym obserwacje nietypowe, a $\{Y_t\}$ będzie nieobserwowanym szeregiem czasowym wolnym od obserwacji nietypowych. W pierwszym etapie [Chen, Liu 1993; Choy 2001] procedury wykrywania obserwacji nietypowych w szeregu czasowym zakłada się stochastyczną strukturę dla nieobserwowanego, wolnego od obserwacji nietypowych szeregu czasowego $\{Y_t\}$ i dokonuje się estymacji parametru modelu, bazując na obserwowanym szeregu czasowym $\{Z_t\}$. Propozycja Foxa została rozszerzona przez różnych autorów [Chan 1995; Tsay 1988] na inne klasy modeli autoregresyjnych, tj. model ze średnią kroczącą – $ARMA(p, q)$ oraz model $ARIMA(p, d, q)$.

W niniejszym opracowaniu zaprezentujemy iteracyjną procedurę wykrywania obserwacji nietypowych opartą na teście współczynnika prawdopodobieństwa, zastosowaną dla modelu $ARIMA(p, d, g)$. W metodzie tej, oprócz zaproponowanych przez Foxa obserwacji nietypowych typu (AO) i (IO), wprowadzono jeszcze dwie kategorie, a mianowicie przesunięcie poziomu (*level shifts* (LS)) oraz tymczasowe zmiany (*transient change* (TC)) [Chen, Liu 1993; Tsay 1988].

Niech Y_t będzie stochastycznym procesem stosowanym w modelu $ARIMA(p, d, g)$:

$$\Phi_p(B)\nabla^d Y_t = \Theta_q(B)\varepsilon_t, \quad (1)$$

gdzie: B – operator przesunięcia wstecz, taki że $BY_t = Y_{t-1}$,

$$\Phi_p(B) = 1 - \varphi_1 B^1 - \varphi_2 B^2 - \dots - \varphi_p B^p \text{ oraz } \Theta_q(B) = 1 - \vartheta_1 B^1 - \vartheta_2 B^2 - \dots - \vartheta_q B^q$$

są wielomianami charakterystycznymi omawianego procesu odpowiednio stopnia p i q ,

Y_t – szereg czasowy wolny od obserwacji nietypowych,

∇^d – operator różnic rzędu d ,

ε_t – zakłócenie losowe (zmienna losowa o rozkładzie normalnym $N(0, \sigma_\varepsilon^2)$).

W celu wykrycia obserwacji nietypowej, zgodnie z wcześniejszym założeniem, zamiast procesu Y_t wykorzystuje się szereg czasowy Z_t zawierający obserwacje nietypowe. Szereg ten dany jest wzorem:

$$Z_t = \omega_i V_i(B) I_t^{(\tau_i)} + Y_t, \quad (2)$$

gdzie: ω_i – reprezentuje wielkość (znaczenie) i -tej obserwacji nietypowej,

$V_i(B)$ – definiuje rodzaj obserwacji nietypowej w następujący sposób:

$$V_i(B) = \begin{cases} 1/\pi(B) & \text{dla (IO)} \\ 1 & \text{dla (AO)} \\ 1/(1-B) & \text{dla (LS)} \\ 1/(1-\delta B) & \text{dla (TC)} \end{cases},$$

gdzie: δ – tempo wygaszania skutków obserwacji nietypowej typu TC $0 < \delta < 1$,

$\pi(B) = \Phi(B) \nabla^d / \Theta(B)$ – autokorelacyjny reprezentant modelu ARIMA (p, d, q) ,

$$I_t^{(\tau_i)} = \begin{cases} 1, & t = T_i \\ 0, & t \neq T_i \end{cases} \text{ – wskaźnik zmiennej reprezentujący obecność lub nieobec-}$$

ność obserwacji nietypowej w czasie T_i .

W celu przetestowania występowania w szeregu czasowym pojedynczej obserwacji nietypowej formuluje się następujące hipotezy badawcze:

$H_0 : \omega_A = \omega_I = \omega_L = \omega_C$ – brak obserwacji nietypowych,

$H_I - \omega_I \neq 0$ – tylko (IO),

$H_A - \omega_A \neq 0$ – tylko (AO),

$H_L - \omega_L \neq 0$ – tylko (LS).

$H_C - \omega_C \neq 0$ – tylko (TC).

Współczynnikami prawdopodobieństwa statystyki testu dla testowanej hipotezy H_0 przeciw H_I , H_A , H_L lub H_C są, kolejno $\lambda_{I,T}$, $\lambda_{A,T}$, $\lambda_{L,T}$ oraz $\lambda_{C,T}$. Zgodnie z hipotezą zerową o braku obserwacji nietypowych, statystyka ta ma asymptotyczny rozkład normalny $N(0; 1)$. Ponieważ parametry modelu wyjściowego są nieznane, to początkowo zostają oszacowane przy założeniu braku obserwacji nietypowych w szeregu czasowym i wówczas wykrywanie wartości zanieczyszczających oparte jest na statystyce $\hat{\lambda}_{I,T}$, $\hat{\lambda}_{A,T}$, $\hat{\lambda}_{L,T}$ oraz $\hat{\lambda}_{C,T}$, gdzie nieznane parametry zastąpiono ich ocenami.

Statystyka testu współczynnika prawdopodobieństwa wyraża się wzorem:

$$\hat{\lambda}_{i,T} = \frac{\hat{\omega}_{i,T}}{\sqrt{\text{var}(\hat{\omega}_{i,T})}}. \quad (3)$$

W celu wykrycia obserwacji nietypowej o nieznannej pozycji, należy w każdym punkcie próbki obliczyć następującą statystykę [Mira, Sanchez 2004]:

$$\eta_t = \max \left\{ \left| \hat{\lambda}_{i,T} \right|, \left| \hat{\lambda}_{A,T} \right|, \left| \hat{\lambda}_{L,T} \right| \right\}. \quad (4)$$

Następnie wartość tej statystyki porównywana jest do z góry ustalonej stałej wartości krytycznej C . Możliwe wartości C , sugerowane przez Chang [Chang, Tiao, Chen 1988], wynoszą: $C = 4,0$ dla szeregów czasowych o niskiej wrażliwości na zdarzenia wywołujące obserwacje nietypowe w szeregach czasowych, $C = 3,5$ dla szeregów o średniej wrażliwości oraz $C = 3,0$ dla szeregów czasowych o wysokiej wrażliwości na zaburzenia zewnętrzne. Jeżeli $\eta_T = \left| \hat{\lambda}_{i,T} \right| \geq C$, wówczas wnioskuje się iż w $t = T$ występuje obserwacja nietypowa o charakterze innowacyjnym (IO), a jeśli $\eta_T = \left| \hat{\lambda}_{A,T} \right| \geq C$, to w okresie $t = T$ występuje addytywna obserwacja nietypowa (AO), jeśli $\eta_T = \left| \hat{\lambda}_{L,T} \right| \geq C$, to przyjmuje się iż w $t = T$ mamy do czynienia z przesunięciem poziomu (LS), jeżeli $\eta_T = \left| \hat{\lambda}_{C,T} \right| \geq C$, to przyjmuje się iż w $t = T$ występuje obserwacja nietypowa określana mianem tymczasowej zmiany (TC).

Inne podejście do wykrywania obserwacji nietypowych w szeregach czasowych związane jest z szacowaniem parametrów modelu metodą odporną (*robust method*) a następnie wykrywanie obserwacji nietypowych przez analizę reszt modelu. W tym celu wykorzystuje się tzw. reszty standaryzowane¹. Estymatory odporne [Zeliaś 1986] charakteryzują się małą wrażliwością na występowanie w danych obserwacji nietypowych, przy czym odporność należy tu rozumieć w zakresie jedynie jako odporność na występowanie obserwacji nietypowych.

W opracowaniu tym skoncentrujemy się jedynie na pierwszym podejściu wykrywania obserwacji nietypowych w szeregach czasowych, bazującym na modelu addytywnych (AO) oraz innowacyjnych (IO) obserwacji nietypowych, a także przesunięciu poziomu (LS) i tymczasowych zmianach (TC)

3. Kriging – metoda statystycznej interpolacji

Kriging jest podstawową metodą estymacji danych wykorzystywaną w geostatystyce. Jest to metoda interpolacji, prognozująca nieznanne wartości zmiennej zre-

¹ O odpornych metodach szacowania parametrów modelu i wykrywaniu na ich podstawie obserwacji nietypowych można przeczytać np.: [Heilpern 1999, s. 312-372].

gionalizowanej² na podstawie jej obserwowanych danych w innych położeniach. W literaturze przedmiotu kriging określany jest mianem najlepszego liniowego nieobciążonego estymatora (BLUE).

W celu wyjaśnienia istoty tej geostatystycznej metody estymacji posłużymy się najprostszą formą krigingu, a mianowicie krigingiem punktowym, który wyraża się wzorem:

$$Z^* = \sum_{i=1}^n \omega_i Z_i, \quad (5)$$

gdzie: $\sum_{i=1}^n \omega_i = 1$,

ω_i – wagi wyznaczone dla każdego z punktów pomiarowych,

Z_i – wartości zmiennej zregionalizowanej w punktach pomiarowych,

n – liczba rozrzuconych punktów pomiarowych w zbiorze.

Podstawą krigingu jest funkcja wariogramu, określana również semiwariogramem lub semiwariancją. Semiwariogram jest funkcją struktury zmiennych zregionalizowanych, która przedstawia przestrzenne lub czasowe zachowanie się tej zmiennej w badanym zbiorze danych. Semiwariancja próbkowa jest definiowana jako połowa wariancji różnicy wartości zmiennych losowych oddalonych o wektor separacji h :

$$\gamma(h) = \frac{1}{2} \text{Var}\{Z(x) - Z(x+h)\}. \quad (6)$$

Estymacja wariogramu sprowadza się do obliczenia empirycznego wariogramu (*experimental variogram*) rozważanej zmiennej zregionalizowanej, a następnie do dopasowania odpowiedniego teoretycznego wariogramu (*model variogram*) do przebiegu semiwariogramu empirycznego.

Wariogram empiryczny charakteryzuje przestrzenną korelację próby losowej. Jest to krzywa (wektor), która powstaje przez graficzne przedstawienie zależności semiwariancji od odległości h między punktami pomiarowymi. Eksperymentalny semiwariogram dla wektora separacji h wyznaczany jest z równania [Ploner, Dutter 2000]:

$$\gamma(h) = \frac{\sum_{i=1}^{N(h)} [Z(x) - Z(x+h)]^2}{2N(h)}, \quad (7)$$

gdzie: $N(h)$ – liczba par punktów odległych o h .

Teoretyczny wariogram charakteryzuje przestrzenną korelację dla populacji źródłowej i jest prostą funkcją matematyczną, modelująca trend w empirycznym wariogramie. Wśród powszechnie wykorzystywanych semiwariogramów teoretycznych, w zależności od sposobu uwzględniania wektora separacji h , wyróżnić można [Ploner, Dutter 2000;

² Zmienna zregionalizowana, jest to zmienna, która lokalnie wykazuje aspekt losowy i ma określony rozkład prawdopodobieństwa, z drugiej strony wykazuje aspekt strukturalny, który odzwierciedla wieloskalowe tendencje zjawiska.

Zawadzki 2002]: model sferyczny, model samorodka, model liniowy, model wykładniczy oraz model gaussowski. Modele te określane są mianem modeli podstawowych.

Model samorodka:

$$\gamma(h) = \begin{cases} 0, & \text{gdy } h = 0 \\ 1, & \text{gdy } h \neq 0 \end{cases}$$

Model sferyczny:

$$\gamma(h) = \begin{cases} c_0 + c_1 \left(\frac{3}{2} \frac{h}{a} - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right), & \text{gdy } |h| \leq a \\ c_0 + c_1, & \text{gdy } |h| > a \end{cases}$$

Model wykładniczy:

$$\gamma(h) = c_0 + c_1 \left(1 - e^{-\frac{h}{a}} \right).$$

Model gaussowski:

$$\gamma(h) = c_0 + c_1 \left(1 - e^{-\frac{h^2}{a^2}} \right).$$

Model liniowy:

$$\gamma(h) = c_0 + bh,$$

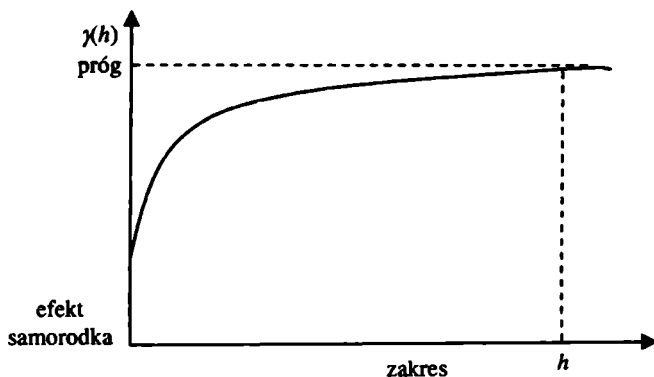
gdzie we wszystkich tych modelach:

c_0 – efekt samorodka (*nugget effect*); c_1 – próg (*sill*).

a – zakres (*range*); b – stosunek progu (c_1) do zakresu (a).

Wymienione elementy: efekt samorodka, próg i zakres są trzema charakterystycznymi parametrami semiwariogramu.

Próg jest maksymalną wartością wariogramu, przy której nie obserwuje się dalszego wzrostu funkcji. Próg powinien w przybliżeniu równać się wariancji próby.



Rys. 1. Najważniejsze charakterystyki semiwariogramu
Źródło: opracowanie własne.

Efekt samorodka jest stałą, która reprezentuje w modelu wariogramu pewną zmienność danych przy skali mniejszej niż przedział próbkowania lub też może być wywołany błędem próbkowania (niską dokładnością pomiaru). Wielkość ta jest pojmowana w semiwariogramie jako wyraz wolny. Zakres jest wielkością skalarną, która kontroluje stopień korelacji między obserwacjami (punktami), zwykle reprezentowanymi jako odległość. Jest to odległość od zera do punktu, w którym semiwariogram osiągnie 95% wartości stałej, będącej progiem (rys. 1).

Zbudowane empiryczne oraz teoretyczne wariogramy mogą być wykorzystane do wyznaczenia wag krigingu ω_i . Na przykład, jeśli chcemy dokonać interpolacji w punkcie P na podstawie danych z trzech otaczających punktów P_1 , P_2 oraz P_3 , równanie krigingu zapisujemy jako:

$$Z_p^* = \omega_1 Z_1 + \omega_2 Z_2 + \omega_3 Z_3. \quad (8)$$

Jeśli używamy semiwariogramu oraz korzystamy z warunku, iż suma wag krigingu powinna być równa jeden, możemy utworzyć następujący układ równań w celu obliczenia wag ω_i :

$$\begin{aligned} \omega_1 \gamma(h_{11}) + \omega_2 \gamma(h_{12}) + \omega_3 \gamma(h_{13}) + \lambda &= \gamma(h_{1p}), \\ \omega_1 \gamma(h_{21}) + \omega_2 \gamma(h_{22}) + \omega_3 \gamma(h_{23}) + \lambda &= \gamma(h_{2p}), \\ \omega_1 \gamma(h_{31}) + \omega_2 \gamma(h_{32}) + \omega_3 \gamma(h_{33}) + \lambda &= \gamma(h_{3p}), \\ \omega_1 + \omega_2 + \omega_3 + 0 &= 1, \end{aligned} \quad (9)$$

gdzie: $\gamma(h_{ij})$ – semiwariancja odległości h między punktami kontrolnymi i oraz j ,

$\gamma(h_{ip})$ – semiwariancja odległości h między punktem kontrolnym i oraz punktem interpolowanym p ,

λ – wyraz wolny równania.

Wartość badanego zjawiska w punkcie P otrzymujemy, rozwiązując układ równań (9) względem wag ω_i , a następnie wagi podstawiamy do równania (8).

Ważną zaletą krigingu jest to, że wariogram może być użyty do wyznaczenia błędu estymacji w każdym punkcie interpolacji. Oszacowanie wariancji można wyznaczyć ze wzoru:

$$S_e^2 = \omega_1 \gamma(h_{1p}) + \omega_2 \gamma(h_{2p}) + \omega_3 \gamma(h_{3p}) + \lambda. \quad (10)$$

4. Zastosowanie metody krigingu w procedurze wykrywania obserwacji nietypowych

Zaprezentowana w niniejszym opracowaniu procedura wykrywania obserwacji nietypowych zakłada, iż w celu wykrycia obserwacji nietypowej o nieznannej pozycji należy w każdym punkcie próbki obliczyć statystkę η . Gdy analizujemy szereg

czasowy składający się z dużej liczby obserwacji i nie znamy synchronizacji (pozycji) oraz typu obserwacji nietypowej, daną statystkę obliczamy dla wszystkich typów obserwacji nietypowych oraz wszystkich danych. To sprawia, że proces wykrywania obserwacji nietypowych staje się niezwykle czasochłonny. W celu wyeliminowania tej niedogodności proponowane są dwa rozwiązania: zastosowanie metody Monte Carlo lub metody krigingu.

Rozwiązaniem tego typu sytuacji, najczęściej wykorzystywanym i prezentowanym w literaturze, jest zastosowanie metod Monte Carlo [Chan 1995] w celu określenia wartości statystyk $\hat{\lambda}_{i,T}$. Mira i Sanchez [2004] proponują w takiej sytuacji wykorzystać metodę krigingu. Postępowanie w tej metodzie jest dwuetapowe. W pierwszym etapie należy wyznaczyć wartości statystyki testu współczynnika prawdopodobieństwa $\hat{\lambda}_{i,T}$ dla wybranych obserwacji szeregu czasowego, a następnie zastosować interpolator krigingu do określenia wartości tej statystyki w innych punktach. W celu wybrania najlepszych „obserwacji reprezentantów”, dla których wartość statystyki będzie wyznaczona zgodnie ze standardową procedurą, autorzy proponują zastosowanie metody analizy czynnikowej lub maksymalnej entropii, albowiem obie metody dają podobne wyniki.

Porównując wartości statystyk otrzymywanych metodami Monte Carlo z tymi otrzymanymi z wykorzystaniem geostatystycznej metody krigingu zaproponowanej przez Mira i Sancheza, należy stwierdzić, że są one na podobnym poziomie, przy czym czasochłonność obliczeń jest nieporównywalnie krótsza w przypadku tej drugiej metody.

5. Wnioski

Wykrywanie obserwacji nietypowych w szeregach czasowych jest zagadnieniem niezwykle ważnym. Występowanie tego typu obserwacji w szeregu czasowym czyni oceny parametrów modeli autoregresyjnych zbudowanych na podstawie takich danych obciążonymi, co powoduje, iż wyniki analiz przeprowadzonych na podstawie tych modeli są nierzetelne oraz zawodne.

Koncepcję wykorzystania metody krigingu w iteracyjnej procedurze wykrywania obserwacji nietypowych jako alternatywnego narzędzia do metod Monte Carlo należy uznać za ciekawą i przynoszącą wiele korzyści. Najważniejszą z nich jest skrócenie czasu obliczeń do kilkudziesięciu sekund, podczas gdy w metodach symulacyjnych czas ten mierzony jest w godzinach. Dodatkową zaletą krigingu jest to, iż dostarcza wariancji oszacowania dla każdego z punktów interpolacji. Ponadto wyniki otrzymywane metodą krigingu są na porównywalnym poziomie do tych otrzymywanych metodami Monte Carlo.

Literatura

- Caroni C., Karioti V., *Detecting an Innovative Outlier in a Set of Time Series*, „Computational Statistics & Data Analysis” 2004, nr 46, s. 561-570.
- Chan W., *Outliers and Financial Time Series Modelling*, „Mathematics and Computers in Simulation” 1995, nr 39, s. 424-430.
- Chang I., Tiao G.C., Chen C., *Estimation of Time Series Parameters in the Presence of Outliers*, „Technometrics” 1988, nr 30, s.190-204.
- Chen C., Liu L., *Joint Estimation of Model Parameters and Outlier Effects in Time Series*, „Journal of the American Statistical Association” 1993, nr 88 (421), s. 284-297.
- Choy K., *Outlier Detection for Stationary Time Series*, „Journal of Statistical Planning and Inference” 2001, nr 99, s. 111-127.
- Fox A.J., *Outliers in Time Series*, „Journal of the Royal Statistical Society, Series B” 1972, nr 34, s. 350-363.
- Heilpern S., *Modelowanie odporne*, [w:] *Statystyczne metody analizy danych*, red. W. Ostasiewicz, AE, Wrocław 1999, s. 312- 372.
- Mira J., Sanchez, M.J., *Prediction of Deterministic Functions, an Application of a Gaussian Kriging Model to a Time Series Outlier Problem*, „Computational Statistics & Data Analysis” 2004, nr 44, s. 477-491.
- Ploner A., Dutter R., *New Directions in Geostatistics*, „Journal of Statistical Planning and Inference” 2000, nr 91, s. 499-509.
- Tsay R.S., *Outliers, Level Shifts, and Variance Changes in Time Series*, „Journal of the American Statistical Association” 1988, nr 81, s. 132-141.
- Urowicz B., *Zastosowanie analizy geostatystycznej I teorii fraktali w badaniach dynamiki wilgotności w profilu glebowym na polach uprawnych*, „Acta Agrophysica” 1999, nr 22, s. 229-243.
- Zawadzki J., *Zastosowanie metod geostatystycznych do analizy danych przestrzennych*, „Wiadomości Statystyczne” 2002, nr 12.
- Zeliaś A., *Metody wykrywania obserwacji nietypowych w badaniach ekonomicznych*, „Wiadomości Statystyczne” 1986, nr 8, s. 16-27.

THE ANALYSIS OF TIME SERIES WITH OUTLIERS USING THE GEOSTATISTICS METHODS

Summary

As a result of action of different external factors, such as: political and economic crises, outbreaks of wars, some observations clearly different from the others, called the outliers, appear in the economic time series.

The outliers generally cause the results of conducted analyses, based on time series, are failed, unreal or even invalid. Therefore, it is important to use some appropriate procedures for detecting and eliminating those effects from time series.

This paper shows the geostatistics kriging method in a detection problem of outliers in time series.