

Alicja Panak

Uniwersytet Łódzki

AUTOKORELACJA PRZESTRZENNA I KRIGING – METODOLOGIA I WYBRANE ZASTOSOWANIA

1. Wstęp

W ostatnich latach miał miejsce gwałtowny rozwój zastosowań modeli przestrzennych w naukach społecznych, biologii i ochronie środowiska, ekonomii i wielu innych. Metodologia ekonometrii przestrzennej [Paelinck, Klaassen 1979; Anselin 1988] ewoluowała od prostych metod statystyki przestrzennej w naukę o regionach i topografii analitycznej [Ord 1975; Cliff, Ord 1981], w formalną gałąź teorii i zastosowań ekonometrycznych [Anselin 1988; Zeliaś i in. 1991; LeSage 1998]. Testowanie autokorelacji przestrzennej oraz estymacja modeli z zależnościami przestrzennymi są coraz powszechniejsze. Techniki regresji przestrzennej są coraz częściej implementowane w narzędziach ekonometrii stosowanej (MATLAB Spatial Econometric toolbox [LeSage 1998], GeoDa [Anselin 2003], R-toolbox).

W opracowaniu prezentowany jest przegląd wybranych publikacji dotyczących opisu i interpretacji przestrzennych modeli regresji. Wyróżniono cztery podstawowe elementy ekonometrii przestrzennej: testowanie występowania i rodzaju związków przestrzennych, konstrukcję schematu ogólnego modelu z uwzględnieniem zależności przestrzennych, estymację parametrów oraz prognozowanie z uwzględnieniem zależności przestrzennych – kriging.

W punkcie drugim zawarto opis modeli autoregresyjnych (z opóźnieniami przestrzennymi) oraz modeli z autokorelowanym składnikiem losowym. W dalszej części zaprezentowano metody testowania samozależności przestrzennych zmiennej prognozowanej Y . Część trzecia opracowania to opis metod estymacji parametrów modeli przestrzennych z użyciem metody największej wiarygodności. W ostatnim punkcie zawarto ogólny schemat metody predykcji uwzględniający zależ-

ności przestrzenne, tzw. kriging klasyczny, wraz ze wskazaniem innych, bardziej złożonych metod prognozowania opartych na teorii procesów losowych.

2. Autokorelacja przestrzenna – przestrzenne modele autoregresyjne

Zwyczajowo rozróżnia się tzw. modele opóźnień przestrzennych (przestrzennej autoregresji) oraz modele przestrzennych składników losowych [Anselin 1988]. Te pierwsze zawierają po prawej stronie równania przestrzennie opóźnioną zmienną zależną WY . Modele przestrzennych składników losowych dotyczą zaś przestrzennej autokorelacji składnika losowego i są modelowane z wykorzystaniem procesu autoregresji składników losowych.

Z metodologicznego punktu widzenia zależności przestrzenne mogą wydawać się dość problematyczne. Zauważmy, że te same zależności mogą stanowić dodatkowe źródło informacji o badanej zmiennej – część wariancji objaśniana jest przez wartość sąsiednią. Uwzględnianie efektów przestrzennych w równaniach ekonometrycznych pociąga za sobą jednak potrzebę specyfikacji przestrzennych zależności w modelu ekonomicznym. Poniżej przedstawione zostały cztery modele przestrzenne: model autoregresyjny I rzędu, model mieszany regresyjno-autoregresyjny, model regresji z autokorelowanym składnikiem losowym oraz ogólny model przestrzenny będący zarówno modelem regresyjno-autoregresyjnym, jak i modelem z autokorelacją składnika losowego. Poniższa prezentacja ogólnych schematów tych modeli ma na celu ułatwienie dalszych rozważań.

W najprostszym przypadku zależności przestrzenne dotyczą opóźnień przestrzennych prognozowanej zmiennej. Rozpatrzmy prosty model autoregresyjny I rzędu – AR(1):

$$Y = \rho WY + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I). \quad (1)$$

Przez Y rozumiemy będziemy n -wymiarowy wektor obserwacji, ρ natomiast jest współczynnikiem autoregresji przestrzennej. ε to n -wymiarowy wektor składników losowych o rozkładzie normalnym z zerową wartością oczekiwaną i wariancją równą $\sigma^2 I$. Macierz reprezentująca przestrzenny charakter zmiennych to tzw. macierz wag W ($n \times n$). Umownie, główną przekątną macierzy W tworzą same zera, pozostałe elementy macierzy – w_{ij} , reprezentują zakładane przestrzenne zależności pomiędzy obserwacjami i i j .

Jeśli w klasycznym modelu regresji uwzględnione zostaną zależności przestrzenne w postaci przestrzennych opóźnień zmiennej zależnej, możemy wówczas rozważać tzw. model mieszany regresyjno-autoregresyjny:

$$Y = \rho WY + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \quad (2)$$

gdzie dodatkowo X jest macierzą zmiennych objaśniających ($n \times k$), a β wektorem nieznanymi współczynników regresji ($k \times 1$).

Wśród modeli przestrzennych składników losowych wskazać możemy klasyczny model regresji z autokorelowanym składnikiem losowym:

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} = \lambda \mathbf{W} \boldsymbol{\varepsilon} + \mathbf{u}, \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (3)$$

w którym $\boldsymbol{\varepsilon}$ to n -wymiarowy wektor skorelowanych składników losowych, $\mathbf{u}$ zaś to wektor składników losowych o rozkładzie normalnym z zerową wartością oczekiwaną i wariancją równą $\sigma^2 \mathbf{I}$. Parametr λ jest współczynnikiem przestrzennej korelacji reszt.

Ogólny model przestrzenny łączy w sobie dwa powyższe modele. Mamy tu do czynienia zarówno z regresją, jak i z autoregresją zmiennej prognozowanej oraz z autokorelacją składnika losowego:

$$\mathbf{Y} = \rho \mathbf{W}_1 \mathbf{Y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} = \lambda \mathbf{W}_2 \boldsymbol{\varepsilon} + \mathbf{u}, \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}). \quad (4)$$

Poniżej przedstawione zostały pewne metody testowania autokorelacji przestrzennej.

Statystyka Morana I służy do testowania obecności autokorelacji przestrzennej według schematu opisanego macierzą wag $\mathbf{W}$. Statystykę I opisuje poniższy wzór:

$$I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (5)$$

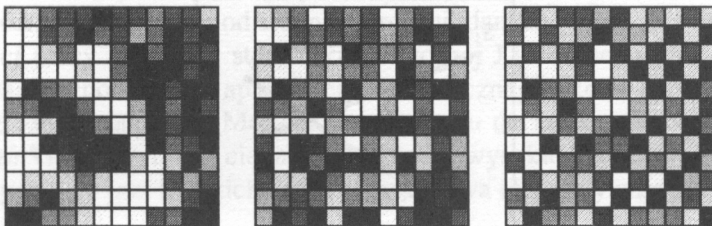
gdzie n – liczba obserwacji, a x_i, x_j – obserwacje o lokalizacji i i j .

Rozpatrzmy hipotezę H_0 : *brak autokorelacji*, przy alternatywnej H_1 : *występowanie autokorelacji*. W przypadku braku autokorelacji statystyka I ma tendencję do przyjmowania wartości bliskiej $-\frac{1}{n-1}$. W przeciwnym wypadku mamy do czynienia z autoko-

relacją dodatnią, gdy $I > -\frac{1}{n-1}$, lub ujemną dla $I < -\frac{1}{n-1}$. Przez autokorelację do-

datnią rozumieć będziemy przestrzenne gromadzenie się (w sensie lokalizacji) wysokich bądź niskich wartości obserwowanych zmiennych. Autokorelację ujemną można interpretować jako odwrotność autokorelacji dodatniej – wysokie wartości sąsiadują z niskimi, i odwrotnie, układając się w pewnego rodzaju szachownicę (rys. 1).

Poprzez sumowanie korelacji pomiędzy wszystkimi punktami i mnożenie ich przez odpowiednie wagi statystyka Morana pokazuje niejako trafność doboru macierzy $\mathbf{W}$. Jeżeli macierz $\mathbf{W}$ odpowiada rzeczywistości, czyli duża waga odpowiada dużej rzeczywistej korelacji, to wartość statystyki Morana jest duża co do modułu.



Rys. 1. Rodzaje autokorelacji przestrzennej

Źródło: opracowanie własne.

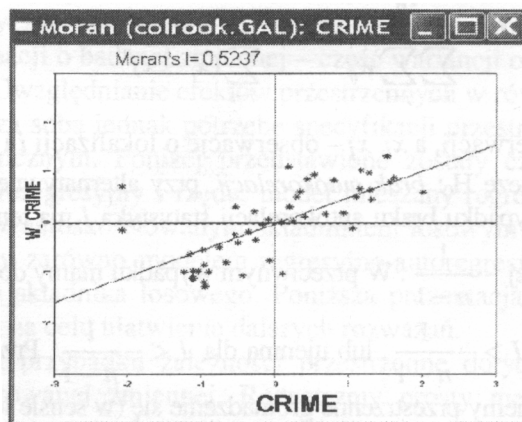
Zauważmy, że dla dostatecznie dużych n w przypadku braku autokorelacji statystyka I przyjmować będzie wartości bliskie zero. W praktyce wartość statystyki Morana rzadko kiedy przekracza co do modułu wartość 1 [Overmars i in. 2003].

Wykres rozproszenia przedstawia związek pomiędzy wartościami:

$(x_i - \bar{x})$ a średnią ważoną sąsiedztwa – $\sum_{i=1}^n w_{ij}(x_i - \bar{x})$, dla wszystkich i . Zgru-

powania w ćwiartkach nieparzystych wskazują na występowanie autokorelacji dodatniej, w parzystych – autokorelacji ujemnej.

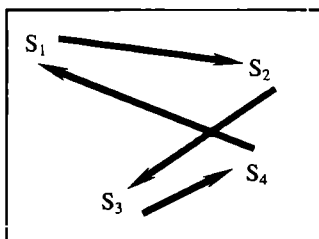
Na rysunku 2 przedstawiono przykład wykresu rozproszenia zrealizowany w programie GeoDa. Wraz punktami rozproszenia na wykresie widoczna jest prosta aproksymująca linię trendu uzyskanego zbioru punktów. Warto zauważyć, iż tangens kąta nachylenia tej prostej jest równy wartości statystyki Morana I .



Rys. 2. Wykres rozproszenia

Źródło: opracowanie własne w programie GeoDa.

W swojej monografii L. Anselin [2003] zaproponował sposób wykorzystania statystyki Morana bazujący na podejściu permutacyjnym. Polega ono na permutowaniu wartości obserwacji przy stałych lokalizacjach. Rysunek 3 przedstawia przykład takiej permutacji.

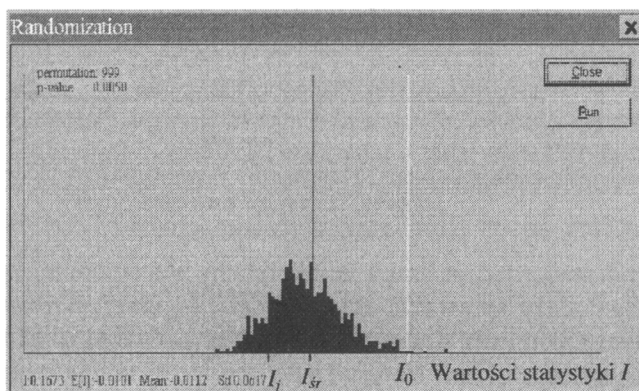


Rys. 3. Metoda permutacyjna

Źródło: opracowanie własne.

Algorytm testu polega na obliczeniu wartości statystyki Morana dla pewnej liczby losowych permutacji zbioru obserwacji i sporządzeniu histogramu testu randomizacyjnego (rys. 4). Jeżeli autokorelacja rzeczywiście występuje, uzyskane w ten sposób wartości statystyki I powinny być co do modułu mniejsze niż w przypadku oryginalnego zbioru danych. Analizując wykres, ustala się, jaki procent permutacji daje większą co do modułu wartość statystyki I . Na tej podstawie można wnioskować o istnieniu bądź braku autokorelacji.

Liczba
permutacji, dla
ustalonych I



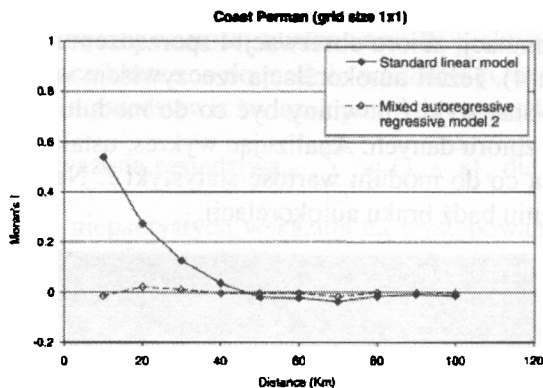
Rys. 4. Test permutacyjny

Źródło: opracowanie własne w programie GeoDa.

Widoczna na rys. 4 wartość p jest pseudopoziomem istotności. Jest to stosunek liczby tych permutacji, dla których $I_i > I_0$ do liczby wszystkich użytych permutacji powiększonej o 1. Zauważmy, że im większa jest wartość p , tym mniej prawdopodobne jest rzeczywiste występowanie autokorelacji.

Statystyka Morana jest najczęściej używaną miarą globalnej autokorelacji przestrzennej, wykorzystywaną w bardzo różnych dziedzinach nauki i tak na przykład K.P. Overmars i inni w swojej pracy [2003] wykorzystują statystykę Morana do wykrywania zależności przestrzennych pomiędzy danymi dotyczącymi zagospodarowania przestrzennego w Ekwadorze. Dzięki tej statystyce udało się nie tylko odnaleźć istniejące zależności przestrzenne, ale i również określić ich siłę (rys. 5). Wykryto dodatnią autokorelację przestrzenną, a także ustalono, że siła korelacji może być różna w zależności od stopnia agregacji danych.

W swojej pracy z zakresu statystyki medycznej J.D. Joines i inni [2003] również korzystali z podobnego aparatu ekonometrycznego. Rozważany tam korelogram uwzględniał statystykę Morana I w stosunku do rzędu sąsiedztwa macierzy oddziaływań W . Odkryli oni ciekawą zależność: wysokie skorelowanie dotyczyło zarówno wysokich, jak i niskich rzędów sąsiedztwa macierzy wag, ale nie pośrednich.

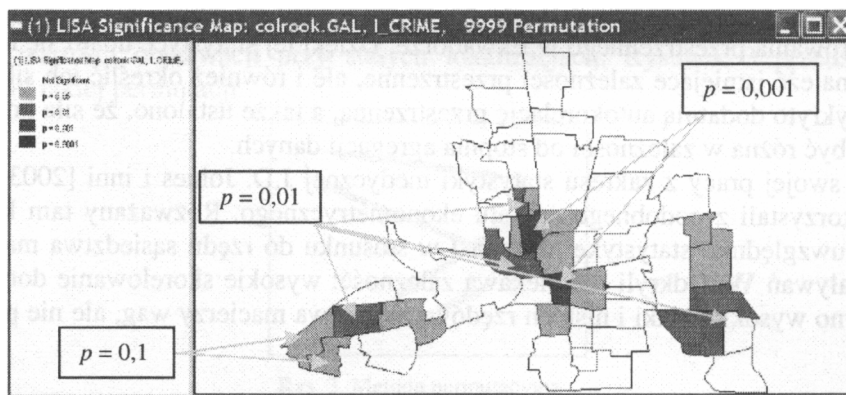


Rys. 5. Porównanie autokorelacji przestrzennej reszt w standardowym modelu liniowym i mieszanym modelu regresyjno-autoregresyjnym
Źródło: [Overmars 2003].

Pomimo stwierdzenia braku autokorelacji przestrzennej zgodnej z zadaną macierzą wag należy sprawdzić istnienie tzw. *lokalnej autokorelacji przestrzennej*, która dotyczy korelacji wartości zmiennej w wybranej lokalizacji z jej sąsiadami. Zadanie to spełnia lokalna statystyka Morana I_i , wyrażająca się wzorem:

$$I_i = \frac{(x_i - \bar{x})}{\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2} \sum_{j=1}^n w_{ij} (x_j - \bar{x}) . \quad (6)$$

Obliczenie statystyki I_i dla wszystkich obserwacji umożliwia sporządzenie tzw. *mapy istotności*, przedstawiającej wartość pseudowspółczynnika istotności testu randomizacyjnego wykonanego dla poszczególnych lokalizacji. Na rysunku 6 przedstawiono przykład mapy istotności zrealizowanej w programie GeoDa. Polom oznaczonym najciemniejszą barwą odpowiada najmniejsza wartość współczynnika pseudoistotności, co oznacza, że są one najbardziej skorelowane ze swoimi sąsiadami.



Rys. 6. Mapa istotności
Źródło: opracowanie własne w programie GeoDa.

Rozważania na temat testowania przestrzennych modeli składnika losowego pominięto ze względu na licznie występujące przykłady opracowań testów służących wykrywaniu istnienia autokorrelacji składnika losowego, tj. testu Morana, mnożnika Lagrange'a, testu Walda itp.

3. Estymacja parametrów modeli przestrzennych – maksymalizacja funkcji gęstości

Ogólnie do estymacji modeli zależności przestrzennych wykorzystuje się zazwyczaj metodę największej wiarygodności. Dla modelu autoregresyjnego I rzędu: (1) $\mathbf{Y} = \rho \mathbf{W} \mathbf{Y} + \boldsymbol{\varepsilon}$, składnik losowy ma rozkład $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. Przypomnijmy, że gdy zmienna $\boldsymbol{\varepsilon}$ ma rozkład o gęstości f oraz istnieje taka funkcja g , przekształcająca zmienną $\boldsymbol{\varepsilon}$ w $\mathbf{Y}$, że $|\det(g^{-1}(\mathbf{y}))| \neq 0$, to zmienna $\mathbf{Y}$ ma rozkład postaci: $\mathbf{Y} \sim f(g^{-1}(\mathbf{Y})) |\det(g^{-1}(\mathbf{Y}))|$. Po dokonaniu pewnych przekształceń, zlogarytmowaniu stronami można otrzymać zlogarytmowaną funkcję wiarygodności L :

$$\ln(L) = -\frac{n}{2} \ln((\mathbf{Y} - \rho \mathbf{W} \mathbf{Y})^T (\mathbf{Y} - \rho \mathbf{W} \mathbf{Y})) + \ln |\det(\mathbf{I} - \rho \mathbf{W})|. \quad (7)$$

Następnie maksymalizuje się funkcję L ze względu na poszukiwany parametr ρ .

W przypadku modelu mieszanego regresyjno-autoregresyjnego: (2) $\mathbf{Y} = \rho \mathbf{W} \mathbf{Y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$ wyznaczenie funkcji wiarygodności jest bardziej kłopotliwe. Pamiętając, że jeżeli

$\mathbf{X} \sim N(\mathbf{b}, \sigma^2 \mathbf{I})$, to $\mathbf{A} \mathbf{X} \sim N(\mathbf{A} \mathbf{b}, \sigma^2 \mathbf{A} \mathbf{A}^T)$, dla $\mathbf{A}$ – będącej macierzą nieosobliwą, można otrzymać warunkowy rozkład funkcji gęstości zmiennej $\mathbf{Y}$: $\mathbf{Y} | \mathbf{X} \sim N((\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta}, \sigma^2 (\mathbf{I} - \rho \mathbf{W})^{-1} (\mathbf{I} - \rho \mathbf{W}^T)^{-1})$.

Dalej można, jak powyżej, skorzystać z metody największej wiarygodności dla parametru ρ oraz wektora parametrów $\boldsymbol{\beta}$ dla funkcji:

$$\begin{aligned} \ln(L_y(\rho, \boldsymbol{\beta})) = & -\ln \sqrt{|\det(\sigma^2 (\mathbf{I} - \rho \mathbf{W})^{-1} (\mathbf{I} - \rho \mathbf{W}^T)^{-1})|} - \\ & - \frac{1}{2} (y - (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta})^T \sigma^2 (\mathbf{I} - \rho \mathbf{W})^{-1} \cdot \\ & (\mathbf{I} - \rho \mathbf{W}^T)^{-1} (y - (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta}). \end{aligned} \quad (8)$$

Dla regresji z autoskorelowanym składnikiem losowym (3) $\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{u}$, $\mathbf{u} = \lambda \mathbf{W} \mathbf{u} + \boldsymbol{\varepsilon}$, tak jak w przypadku modelu AR(1) wywnioskować można, że zmienna $\mathbf{u}$ ma następujący rozkład: $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 (\mathbf{I} - \lambda \mathbf{W})^{-1} (\mathbf{I} - \lambda \mathbf{W}^T)^{-1})$. Zatem dla $\mathbf{V} = \sigma^2 (\mathbf{I} - \lambda \mathbf{W})^{-1} (\mathbf{I} - \lambda \mathbf{W}^T)^{-1}$ funkcja wiarygodności przyjmuje postać:

$$\ln L(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = -\frac{n}{2} \ln |\det \mathbf{V}(\lambda)| - \frac{1}{2} [(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\lambda)^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})]. \quad (9)$$

W przypadku modelu mieszanego regresyjno-autoregresyjnego z autokorelowanym składnikiem losowym przy estymacji parametrów buduje się funkcję wiarygodności, korzystając z poprzednich dwóch modeli.

K.P. Overmars i inni w pracy z 2003 r. opisali swoje badania, w których wykryli istnienie autokorelacji reszt modelu liniowego. Pokazano, że użycie tych samych danych z zastosowaniem metod przestrzennych i klasycznych prowadzi do zupełnie różnych wniosków. W standardowym modelu liniowym reszty okazywały się skorelowane – zwłaszcza dla niskich poziomów agregacji. Aby uniknąć tego problemu, zastosowano mieszany model regresyjno-autoregresyjny z autokorelowanym składnikiem losowym – model (4). Reszty nie wykazywały skorelowania, przy równoczesnej najlepszej „dobroci dopasowania” modelu. We wnioskach autorzy postulują, aby wykrywanie autokorelacji oraz szacowanie skali jej występowania w przypadku zagospodarowywania terenów, których dotyczyła praca, stało się standardową procedurą.

W pracy [Joines 2003] zastosowano modele autoregresyjne dla różnych macierzy wag odpowiadających różnym rzędom ciągłości. Autorzy wskazują, jak duże znaczenie ma uwzględnienie autokorelacji przestrzennej w badaniach dotyczących małych obszarów. Uzyskane w ten sposób informacje pomogły stworzyć model o lepszym dopasowaniu i mniejszych błędach parametrów modelu (ok. 20%) w stosunku do tych otrzymanych klasyczną metodą MNK.

Literatura

- Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht, s. 284.
- Anselin L. (2002), *Under The Hood: Issues in the Specification and Interpretation of Spatial Regression Models*, „Agricultural Economics” 27, s. 247-267.
- Anselin L. (2003), *GeoDa 0.9 User's Guide*, Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, University of Illinois.
- Bandemer H., Gebhardt A. (2000), *Bayesian Fuzzy Kriging*, „Fuzzy Sets and Systems”, vol. 112, s. 405-418.
- Boots B., Tiefelsdorf M. (2000), *Global and Local Autocorrelation in Bounded Regular Tessellations*, „Geographical Systems” 2, s. 319-348.
- Cliff A., Ord J.K. (1981), *Spatial Processes: Models and Applications*, Pion, London.
- Griffith D.A. (2000), *A Linear Regression Solution to the Spatial Autocorrelation Problem*, „Geographical Systems”, s. 141-156.
- Joines J.D., Hertz-Picciotto I., Carey T.S., Gesler W., Suchindran Ch. (2003), *A Spatial Analysis of County-Level Variation in Hospitalization Rates for Low*

Back Problems in North Carolina, „Social Science & Medicine” 56, s. 2541-2553.

LeSage J.P. (1998), *Spatial Econometrics*, Department of Economics, University of Toledo.

Ord, J.K. (1975), *Estimation Methods for Models of Spatial Interaction*, „J. Am. Stat. Assoc.”, 70, s. 120-126.

Overmars K.P., de Koning G.H.J., Veldkamp A. (2003), *Spatial Autocorrelation in Multi-Scale Land Use Models*, „Ecological Modelling”, 164, s. 257-270.

Paelinck J.H.P, Klaassen L.H. (1979), *Spatial Econometrics*, Saxon House.

Zeliaś A., Grabiński T., Ludwiczak B., Malina A. (1991), *Ekonometria przestrzenna*, PWE, Warszawa.

SPATIAL AUTOCORRELATION AND KRIGING – METHODOLOGY WITH SOME APPLICATIONS

Summary

Recent research in spatial econometrics focus on a problem of analysing spatial interactions and spatial autocorrelation. Spatial relations are also extensively used in a geostatistical technique for Gaussian interpolation – kriging. The aim of this work is to present some testing methods for both global and local spatial autocorrelation, and to discuss some estimation and interpolation problems in spatial interferences context. Some applied examples are presented.