Pietrzak, Michal Bernard; Wilk, Justyna; Bivand, Roger; Kossowski, Tomasz

Working Paper

Wpływ wyboru metody klasyfikacji na identyfikację zależności przestrzennych – zastosowanie testu join-count

Institute of Economic Research Working Papers, No. 21/2014

Provided in Cooperation with:
Institute of Economic Research (IER), Toruń (Poland)

Suggested Citation: Pietrzak, Michal Bernard; Wilk, Justyna; Bivand, Roger; Kossowski, Tomasz (2014) : Wpływ wyboru metody klasyfikacji na identyfikację zależności przestrzennych – zastosowanie testu join-count, Institute of Economic Research Working Papers, No. 21/2014, Institute of Economic Research (IER), Toruń

This Version is available at:
https://hdl.handle.net/10419/219582

Institute of Economic Research Working Papers

No. 21/2014

Wpływ wyboru metody klasyfikacji na identyfikację zależności przestrzennych – zastosowanie testu join-count

Michał Bernard Pietrzak

Justyna Wilk

Roger Bivand

Tomasz Kossowski

Toruń, Poland 2014

Michał Bernard Pietrzak
Michal.Pietrzak@umk.pl
Nicolaus Copernicus University in Toruń, Department of Econometrics and Statistics, ul. Gagarina 13a, 87-100 Toruń

Justyna Wilk
justyna.wilk@ue.wroc.pl
Wrocław University of Economics, Department of Econometrics and Computer Science, ul. Nowowiejska 3, 58-500 Jelenia Góra

Roger Bivand
Roger.Bivand@nhh.no
Norwegian School of Economics (NHH) in Bergen, Department of Economics, Helleveien 30, 5045 Bergen

Tomasz Kossowski
tkoss@amu.edu.pl
Adam Mickiewicz University in Poznań, Department of Spatial Econometrics, ul. Dzięgielowa 27, 61-680 Poznań

The influence of classification method selection on the identification of spatial dependence – application of join-count test

JEL Classification: *C21, C51, J64, R11*

Keywords: *join-count test, spatial dependence, classification, qualitative data, economic development.*

Abstract: A lot of regional studies cover classification of territorial units due to considering research problem. This approach examines territorial diversification of a phenomena, as well as spatial interactions. The occurrence of spatial dependence can reveal the processes of creating or extanding spatial clusters. One of the significant determinants of research results is a way of classification. Different approach leads to diversified divisions of territorial units. The objective of this paper is to examine the influence of classification method selection on the spatial autocorrelation analysis results using join-count test. This test, in contradiction to the other methods proposed in the field of spatial statistics, analyzes spatial

autocorrelation based on qualitative data. Therefore, it can be applied in the examination of spatial dependence between territorial units from distinguished classes of territorial units.

Wprowadzenie

Występowanie zależności przestrzennych wskazuje na kształtowanie się zjawisk w odniesieniu do lokalizacji przestrzennej. Zależności przestrzenne stanowią naturalną własność większości zjawisk społeczno-ekonomicznych. Poziom interakcji między jednostkami terytorialnymi jest bowiem tym wyższy, im bliżej są one położone względem siebie. Nieuwzględnienie tych informacji może prowadzić do błędów poznawczych (Zeliaś (red.) 1991, Suchecki (red.) 2010, Arbia 2006).

W przestrzennych badaniach ekonomicznych bardzo często dokonuje się podziału jednostek terytorialnych na kategorie (klasy). W zależności od zastosowanej metody uzyskuje się klasy uporządkowane (np. poziom życia, etap rozwoju społeczno-gospodarczego, poziom innowacyjności, stopień rozwoju turystyki itd.), bądź równorzędne (np. profil gospodarczy, struktura rynku pracy itd.). Zdefiniowane klasy, w odniesieniu do skal pomiaru, stanowią realizacje zmiennych niemetrycznych, tj. porządkowych lub nominalnych (zob. Walesiak 1993).

Pomiar zależności przestrzennych między jednostkami z tej samej klasy pozwala zbadać, czy istnieją mechanizmy wzmacniające proces tworzenia lub rozszerzania się klastrów przestrzennych, np. obszarów metropolitalnych, obszarów zapóźnionych, regionów turystycznych itd. Jednakże w tej sytuacji stosowalność popularnych testów badających autokorelację przestrzenną jest ograniczona.

Celem referatu jest zbadanie wpływu wyboru metody klasyfikacji na wyniki badania zależności przestrzennych. Główna uwaga skierowana została na problem testowania zależności przestrzennych na podstawie danych jakościowych. Zastosowano test join-count do identyfikacji zależności przestrzennych między jednostkami terytorialnymi reprezentującymi podobny poziom rozwoju gospodarczego.

W pierwszej części artykułu omówiono istotę testu join-count. W drugiej części artykułu zaprezentowano przykład empiryczny. Badaniem objęto sytuację 379 powiatów (LAU 1) w 2012 roku. W tym celu skonstruowano taksonomiczny miernik rozwoju (TMR). Na podstawie wartości TMR powiaty przyporządkowane zostały do klas reprezentujących zróżnicowany poziom rozwoju gospodarczego. Następnie za pomocą testu join-count zbadano występowanie autokorelacji przestrzennej.

Istota testu join-count

W identyfikacji zależności przestrzennych w przypadku zjawisk społeczno-ekonomicznych stosowana jest najczęściej funkcja autokorelacji przestrzennej (Chojnicki (red.) 1980, Suchecki (red.) 2010, s. 103-104, Kossowski 2010). Popularnymi testami, pozwalającymi ocenić siłę zależności przestrzennych, są statystyka I Morana oraz statystyka C Geary'ego (Moran 1947, Cliff, Ord 1973, Cliff, Ord 1981, Kopczewska 2006, s. 69-70, Suchecki (red.) 2010, s. 112-115, Suchecka (red.) 2014, s. 41). Są one przeznaczone do analizy danych ilościowych, np. wartości PKB.

W przypadku danych jakościowych, pomiar zależności przestrzennych może być prowadzony z wykorzystaniem testu join-count (zob. Cliff, Ord 1973, Cliff, Ord 1981, Kopczewska 2006, s. 83-84, Suchecki (red.) 2010, s. 110-112, Pietrzak, Wilk, Kossowski, Bivand 2014, Pietrzak, Wilk, Bivand, Kossowski 2014). Rozkład przestrzenny wartości zmiennej niemetrycznej może być losowy, bądź wykazywać tendencję do przestrzennego grupowania się. W przypadku występowania dodatniej autokorelacji przestrzennej, dominować powinno sąsiedztwo jednostek tej samej kategorii nad sąsiedztwem jednostek różnych kategorii. W przeciwnym wypadku można przyjąć występowanie ujemnej autokorelacji.

W najprostszym przypadku przyjmuje się, że zmienna niemetryczna przyjmuje dwie realizacje. Na kartogramie można je zaprezentować za pomocą koloru białego (W – white) i czarnego (B – black). Idea wyznaczania statystyk join-count polega na zliczaniu sąsiedztwa typu białe-białe (WW), czarne-czarne (BB) oraz czarne-białe (BW). Wyznaczane są zatem trzy statystyki (zob. Cliff, Ord 1973, 1981):

$$WW = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(1-z_i)(1-z_j), \qquad (1)$$

$$BB = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j, \qquad (2)$$

$$BW = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - z_j)^2, \qquad (3)$$

gdzie: $WW + BW + BB = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij},$

z_i, z_j – zmienne zero-jedynkowe przyjmujące wartość 1 w sytuacji, gdy region należy do klasy „czarny" (B), natomiast wartość 0, gdy region należy do klasy „biały" (W),

w_{ij} – element przyjętej macierzy sąsiedztwa,

i, j – numery jednostek terytorialnych ($i,j = 1, \ldots, n$).

Jeżeli sąsiedztwa „jednokolorowe" nie dominują wyraźnie nad „dwukolorowymi" (i na odwrót), to oznaczać może to losowy rozkład wartości zmiennej.

Pomiar poziomu rozwoju gospodarczego

W przestrzennych analizach poziomu rozwoju gospodarczego i jego zmian szczególne zastosowanie mają metody taksonomiczne (zob. Chojnicki i Czyż 1973, Grabiński, Wydymus i Zeliaś 1989, Strahl (red.) 2006). Pierwszą grupę tych metod stanowi analiza skupień (zob. np. Everitt, Landau i Leese 2001). Pozwala ona na wyodrębnienie względnie jednorodnych wewnętrznie i separowalnych zewnętrznie klas obiektów. Jest ona przydatna m.in. w sytuacji, gdy celem badania jest porównanie struktury rynku pracy, profilu gospodarczego jednostek terytorialnych itd.

Do drugiej grupy zaliczają się metody porządkowania liniowego (zob. np. Hellwig 1968). Służą one uporządkowaniu obiektów według nadrzędnego kryterium, które nie podlega pomiarowi bezpośredniemu. Pozwalają one określić odległość obiektów od pewnego, najczęściej z góry ustalonego wzorca rozwoju. Są one powszechnie stosowanym narzędziem w pomiarze poziomu rozwoju gospodarczego, a uzyskane wyniki mogą stanowić podstawę do wydzielenia klas regionów reprezentujących zróżnicowany poziom rozwoju gospodarczego.

W artykule przeprowadzono analizę poziomu rozwoju gospodarczego w 379 polskich powiatach w 2012 roku. Pomiar poziomu rozwoju gospodarczego wymagał rozważenia wielu aspektów, takich jak profil gospodarczy i aktywność gospodarcza, produkcyjność i kondycja przemysłu, przedsiębiorczość i skłonność do inwestycji, chłonność rynku pracy i adaptacyjność zasobów pracy, napływ kapitału zagranicznego, a także sytuacja finansowa mieszkańców i ich zdolność nabywcza.

Ze względu na uwzględnienie jednostek terytorialnych szczebla lokalnego wystąpiły problemy z dostępnością danych statystycznych. Niektóre wartości nie są wyznaczane dla powiatów (np. wartość dodana brutto), bądź część danych jest ukryta (np. wartość kapitału zagranicznego).

W takiej sytuacji konstruowano zmienną alternatywną, a jeśli nie było takiej możliwości – eliminowano aspekt z analizy. Zmienne dobrano w taki sposób, aby spełniały kryterium porównywalności, jednoznacznego definiowania problemu, mierzalności i przydatności w opisie zjawisk na poziomie lokalnym (LAU 1) oraz niepowielania informacji i wykazywania statystycznej zmienności. Ostateczny zestaw zmiennych zawarto w tabeli 1.

Tabela 1. Zestaw zmiennych opisujących poziom rozwoju gospodarczego powiatów

Lp.	Nazwa zmiennej	Charakter zmiennej	Wartość wzorca	Wartość antywzorca
1	Nakłady inwestycyjne w przedsiębiorstwach* na 1 mieszkańca w wieku produkcyjnym (średnia w okresie 2010-2012) [zł]	stymulanta	31798,60	362,90
2	Przeciętne miesięczne wynagrodzenie brutto* [zł]	stymulanta	6541,95	2349,11
3	Stopa bezrobocia rejestrowanego [%]	destymulanta	4,20	38,00
4	Podmioty z udziałem kapitału zagranicznego na 10 tys. mieszkańców [jedn. gosp.]	stymulanta	47,85	0,00
5	Osoby fizyczne prowadzące działalność gospodarczą na 100 osób w wieku produkcyjnym [jedn. gosp.]	stymulanta	20,00	5,30
6	Pracujący w handlu i usługach** na 1000 ludności w wieku produkcyjnym [osoba]	stymulanta	247,67	17,26
7	Podmioty gospodarki narodowej nowo zarejestrowane w rejestrze REGON na 10 tys. ludności [jedn. gosp.]	stymulanta	214,00	40,00

* bez podmiotów gospodarczych o liczbie pracujących do 9 osób

** uwzględniono sekcje PKD 2007: handel; naprawa pojazdów samochodowych; transport i gospodarka magazynowa; zakwaterowanie i gastronomia; informacja i komunikacja

Źródło: Opracowanie własne na podstawie danych BDL GUS.

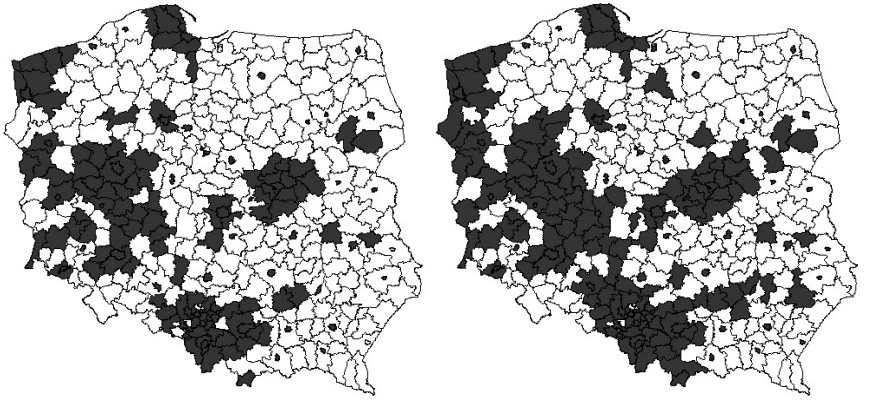
Następnie przystąpiono do konstrukcji taksonomicznego miernika rozwoju. W pierwszym kroku zdefiniowano obiekt wzorzec oraz obiekt antywzorzec. Większość zmiennych (za wyjątkiem stopy bezrobocia rejestrowanego) ma charakter stymulant, dlatego za wartości wzorcowe uznano maksimum dla stymulant i minimum dla destymulant. Współrzędne antywzorca wyznaczono w sposób odwrotny. Następnie zastosowano unitaryzację zerowaną w celu normalizacji wartości zmiennych, a także przekształcono destymulantę na stymulantę poprzez odjęcie jej wartości od jedności. Następnie wyznaczono odległości euklidesowe obiektów (powiatów) od wzorca i antywzorca. Wartości TMR wyznaczono poprzez podzielenie odległości od wzorca przez sumę odległości od wzorca i

antywzorca. Miernik przyjął wartości w przedziale [0, 1], gdzie wartość 1 oznacza wzorzec, a 0 antywzorzec.

Identyfikacja zależności przestrzennych w analizie poziomu rozwoju gospodarczego

W celu porównania sytuacji powiatów, na podstawie wartości miernika TMR, wydzielono klasy reprezentujące zróżnicowany poziom rozwoju gospodarczego. W pierwszym podejściu powiaty podzielono na dwie grupy „W" (biały) oraz „B" (czarny), stosując dwa kryteria podziału: średnią arytmetyczną oraz medianę. Klasa „W" skupia powiaty o relatywnie słabym, natomiast klasa „B" – relatywnie wysokim poziomie rozwoju gospodarczego (zob. rys. 1).

Rysunek 1. Podział powiatów na dwie klasy



a) podział oparty na średniej arytmetycznej b) podział oparty na medianie

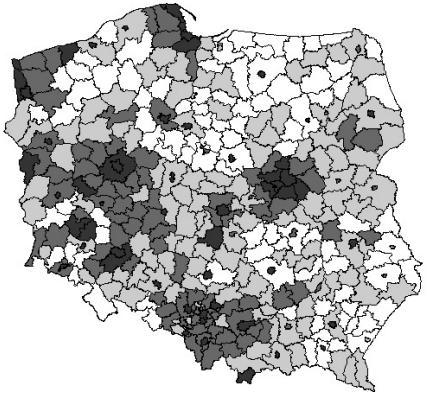
Źródło: Opracowanie własne na podstawie wartości taksonomicznego miernika rozwoju.

Tabela 2. Wyniki testu *join-count* w podziale powiatów na dwie klasy

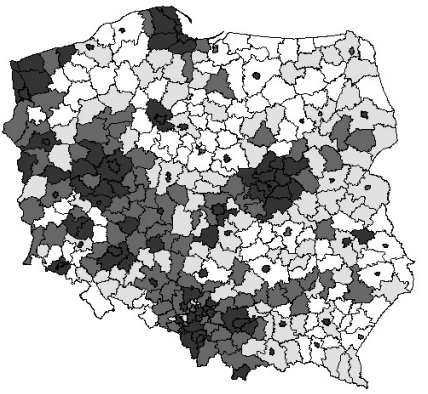
Metoda klasyfikacji	Rodzaj testu	Liczebność klasy	Statystyka	Wartość oczekiwana	Wariancja	Statystyka s.	Wartość p
Średnia arytmetyczna	WW	224	489,92	348,24	197,42	10,08	0,00
	BB	155	219,04	166,73	131,52	4,56	0,00
Mediana	WW	190	381,34	247,66	152,7552	10,82	0,00
	BB	189	302,73	250,86	147,3044	4,27	0,00

Źródło: Opracowanie własne w pakiecie `spdep` (Bivand i in. 2014) programu `R-CRAN`.

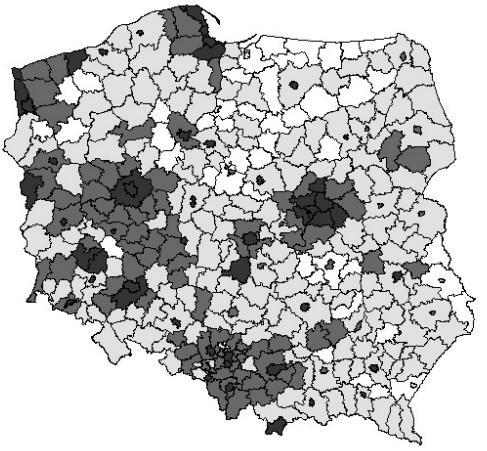
Rysunek 2. Podział powiatów na cztery klasy



a) metoda trzech średnich



b) metoda trzech median



c) podział oparty na średniej arytmetycznej i odchyleniu standardowym

Źródło: Opracowanie własne na podstawie wartości taksonomicznego miernika rozwoju.

Na podstawie przeprowadzonych testów stwierdzono statystyczną istotność zależności przestrzennych zarówno w przypadku przestrzennego skupiania się powiatów z klasy pierwszej (statystyka WW), jak i przestrzennego skupiania się powiatów z klasy drugiej (statystyka BB), w obu podziałach. Potwierdziło to zatem występowanie klastrów przestrzennych, tj. tendencję do przestrzennego skupiania się powiatów o

relatywnie wysokim poziomie rozwoju gospodarczego oraz podobną sytuację w przypadku powiatów słabszych.

W kolejnym podejściu powiaty przyporządkowano do czterech klas: „A", „B", „C", „D", gdzie klasa „A" oznacza bardzo słaby, klasa „B" słaby, klasa „C" umiarkowany, natomiast klasa „D" – relatywnie wysoki poziom rozwoju gospodarczego. W tym celu zastosowano trzy popularne metody klasyfikacji, tj. metodę trzech średnich, metodę trzech median oraz metodę opartą na średniej arytmetycznej i odchyleniu standardowym (zob. rys. 2).

Tabela 3. Wyniki testu *join-count* w podziale powiatów na cztery klasy

Metoda klasyfikacji	Rodzaj testu	Liczebność klasy	Statystyka	Wartość oczekiwana	Wariancja	Statystyka stand.	Wartość p
Metoda trzech średnich	AA	94	121,87	61,02	47,74	8,81	0,00
	BB	130	164,32	116,80	92,07	4,95	0,00
	CC	99	112,62	67,80	60,23	5,78	0,00
	DD	56	23,09	21,42	20,65	0,37	0,36
Metoda trzech median	AA	95	123,47	62,18	48,22	8,83	0,00
	BB	95	95,04	61,00	54,27	4,62	0,00
	CC	93	90,43	61,58	55,07	3,89	0,00
	DD	96	87,99	61,99	54,49	3,52	0,00
Średnia arytmetyczna i odchylenie standardowe	AA	46	17,28	14,02	13,67	0,88	0,19
	BB	48	120,00	14,59	14,03	28,14	0,00
	CC	233	300,00	220,21	147,23	6,58	0,00
	DD	52	40,00	18,56	17,54	6,78	0,00

Źródło: Opracowanie własne w pakiecie `spdep` (Bivand i in. 2014) programu `R-CRAN`.

Testowano występowanie dodatniej autokorelacji przestrzennej w klasach poprzez wyznaczenie statystyk AA, BB, CC, DD. Uzyskane wyniki zaprezentowane zostały w tabeli 3. W tym przypadku występowanie dodatnich zależności przestrzennych we wszystkich klasach potwierdzono tylko w podziale z wykorzystaniem metody trzech median. Natomiast brak statystycznej istotności zależności przestrzennych dotyczył klasy o najwyższym poziomie rozwoju gospodarczego w podziale metodą trzech średnich, a także klasy o najniższym poziomie rozwoju gospodarczego w podziale opartym na średniej arytmetycznej i odchyleniu standardowym.

Podsumowanie

W pracy podjęto próbę zbadania wpływu wyboru metody klasyfikacji na identyfikację zależności przestrzennych z wykorzystaniem testu join-count. W ramach statystyki przestrzennej test ten jest stosowany w analizie

losowości reszt modelu regresji, a także w badaniu autokorelacji przestrzennej na podstawie danych jakościowych.

W badaniu zastosowano łącznie pięć podziałów 379 polskich powiatów, na dwie oraz cztery klasy, reprezentujące zróżnicowany poziom rozwoju gospodarczego w 2012 roku. Przyjęta procedura pozwoliła ujawnić występujące zależności przestrzenne, bądź potwierdzić ich brak. Zależności przestrzenne, cechujące wszystkie klasy, dotyczyły podziału powiatów na dwie grupy z wykorzystaniem średniej arytmetycznej oraz mediany, a także podziału na cztery grupy metodą trzech median.

Metody te wskazały, że w Polsce przestrzennie grupują się jednostki o relatywnie wysokim poziomie rozwoju gospodarczego, szczególnie w ramach tworzenia się obszarów metropolitalnych. Grupują się także jednostki o umiarkowanym poziomie rozwoju, co może wynikać z dyfuzyjnego oddziaływania regionów mocniejszych gospodarczo. Uwidacznia się również tworzenie tzw. klastrów biedy, gdy grupują się przestrzennie obszary o najsłabszym poziomie rozwoju gospodarczego.

W badaniu wykazano związek między metodą klasyfikacji i występowaniem zależności przestrzennych. Proponowana procedura może służyć wyborowi metody klasyfikacji oraz liczby klas w przypadku, gdy badacz poszukuje podziału, w którym jednostki zgrupowane w klasy wykazują dodatnią autokorelację przestrzenną.

Literatura

Arbia G., 2006, *Spatial econometrics*, Springer, Berlin- Heidelberg.

Bivand R. S. (red.), 2014, *spdep package*, R-CRAN, http://cran.r-project.org/web/packages/spdep/index.html.

Bivand R. S., Pebesma E. J., Gómez-Rubio V., 2008, *Applied Spatial Data Analyses with R*, Springer, New York.

Chojnicki Z. (red.), 1980, *Analiza regresji w geografii*, PWN, Warszawa.

Chojnicki Z., Czyż T., 1973, *Metody taksonomii numerycznej w regionalizacji geograficznej*, PWN, Warszawa.

Cliff, A. D., Ord J. K., 1973, *Spatial Autocorrelation,* Pion, London.

Cliff, A. D., Ord J. K., 1981, *Spatial Processes: Models and Applications*, Pion, London.

Everitt B.S., Landau S., Leese M., 2001, *Cluster Analysis*, Fourth Edition, Arnold, London.

Grabiński T., Wydymus S., Zeliaś A., 1989, *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, PWN, Warszawa.

Hellwig Z., 1968, *Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę*

wykwalifikowanych kadr, "Przegląd Statystyczny", R. XV, zeszyt 4, s. 307-327.

Kopczewska K., 2006, *Ekonometria i statystyka przestrzenna z wykorzystaniem programu R CRAN*, Cedetu, Warszawa.

Kossowski T., 2010, *Teoretyczne aspekty modelowania przestrzennego w badaniach regionalnych*, [w:] P. Churski (red.), Praktyczne aspekty badań regionalnych, Biuletyn Instytutu Geografii Społeczno-Ekonomicznej i Gospodarki Przestrzennej Uniwersytetu Adama Mickiewicza w Poznaniu, nr 12, Bogucki Wydawnictwo Naukowe, Poznań.

Moran P. A. P., 1947, *The interpretation of statistical maps*, Journal of the Royal Statistical Society, B10, s. 243-251.

Pietrzak B., Wilk J., Kossowski T., Bivand R., 2014, *The identification of spatial dependence in the analysis of regional economic development - join-count test application*, [w:] M. Papież, S. Śmiech (red.), Proceedings of the 8th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Wyd. Uniwersytetu Ekonomicznego w Krakowie, Kraków, s. 135-144.

Pietrzak B., Wilk J., Bivand R., Kossowski T., 2014, The Application of Local Indicators for Categorical Data (LICD) in Identifying of Spatial Dependences in the Analysis of Socio-Economic Development, "Comparative Economic Research. Central and Eastern Europe.", No. 4.

Strahl D. (red.), 2006, *Metody oceny rozwoju regionalnego*, Wyd. Akademii Ekonomicznej we Wrocławiu, Wrocław.

Suchecka J. (red.), 2014, *Statystyka przestrzenna. Metody analiz struktur przestrzennych*, C.H. Beck., Warszawa.

Suchecki B. (red.), 2010, Ekonometria *przestrzenna. Metody i modele analizy danych przestrzennych*, Wydawnictwo C.H. Beck, Warszawa.

Walesiak M., 1993, *Strategie postępowania w badaniach statystycznych w przypadku zbioru zmiennych mierzonych na skalach różnego typu,* "Badania Operacyjne i Decyzje", nr 1, s. 71-77.

Zeliaś A. (red.),1991, *Ekonometria przestrzenna*, PWE, Warszawa.