PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Elżbieta Antczak, Karolina Lewandowska-Gwarda
Uniwersytet Łódzki

ZASTOSOWANIE METOD EKSPLORACYJNEJ ANALIZY DANYCH PRZESTRZENNYCH W BADANIU POZIOMU UMIERALNOŚCI W POLSCE

1. Wstęp

Eksploracja danych jest międzydyscyplinarną dziedziną nauki, która zajmuje się odkrywaniem wiedzy z danych statystycznych. Są w niej stosowane metody ilościowe właściwe dla poszczególnych rodzajów danych statystycznych, tj. szeregi czasowe, dane panelowe czy przestrzenne, w celu znalezienia współzależności, wzorców, trendów i podsumowania danych w oryginalny sposób. Eksploracyjna analiza danych przestrzennych jest dziedziną wiedzy, która również wykorzystuje metody wizualizacji danych (na mapach i wykresach) w celu wydobycia jak największej ilości informacji z banków danych przestrzennych [Larose 2006, s. XI-2]. Dane przestrzenne są bardziej skomplikowane w swojej strukturze niż szeregi czasowe. Badając obiekty przestrzenne (kraje, regiony, województwa), należy pamiętać, że nie są one izolowane w przestrzeni i mogą podlegać wpływom innych jednostek. Położenie analizowanych obiektów (sąsiedztwo, odległość) wpływa na kształtowanie się interakcji przestrzennych. Zgodnie z prawem analiz przestrzennych sformułowanym przez W.R. Toblera „wszystko jest powiązane ze sobą, ale bliższe obiekty są bardziej zależne od siebie niż odległe" (zob. [Tobler 1970, s. 236]). Konsekwencją tego może być przestrzenne grupowanie się podobnych wartości zmiennych zlokalizowanych (autokorelacja dodatnia) bądź ich dyspersja (autokorelacja ujemna). Ponadto − w odróżnieniu od jednokierunkowych zależności obserwowanych w przypadku szeregów czasowych − dla danych przestrzennych zależności są z reguły wielokierunkowe. Zastosowanie klasycznych metod ilościowych w analizach danych przestrzennych uniemożliwia więc poprawny opis badanego zjawiska.

Celem opracowania jest prezentacja wybranych metod eksploracyjnej analizy danych przestrzennych ESDA (*Exploratory Spatial Data Analysis*), która jest zbiorem technik statystycznych wykorzystywanych do charakterystyki obserwacji wykazujących zależności przestrzenne. W artykule omówione zostały podstawowe narzędzia ESDA:

statystyki lokalnej i globalnej autokorelacji przestrzennej zmiennych. Zastosowania ESDA zaprezentowano na przykładzie dotyczącym umieralności z powodu chorób układu krążenia w Polsce na poziomie powiatów. Analizą objęto rok 2006.

2. Eksploracyjna analiza danych przestrzennych – narzędzia

Podstawowymi narzędziami eksploracyjnej analizy danych przestrzennych są:

- globalna statystyka Morana I − globalny wskaźnik powiązań przestrzennych – która odpowiada na pytanie: czy autokorelacja przestrzenna występuje w badanym obszarze,
- lokalna statystyka Morana I_i − lokalny wskaźnik powiązań przestrzennych LISA (*Local Indicators of Spatial Association)* – która odpowiada na pytanie: gdzie dokładnie (w której części badanego obszaru) występuje autokorelacja przestrzenna.

Wartość globalnej statystyki Morana I dla standaryzowanej przestrzennej macierzy wag **W**, w której elementy w każdym wierszu sumują się do 1, opisuje wzór [Le Gallo, Ertur 2003, s. 179-180]:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathbf{z}^T\mathbf{W}\mathbf{z}}{\mathbf{z}^T\mathbf{z}}. \tag{1}$$

Natomiast dla niestandaryzowanej przestrzennej macierzy wag wartość statystyki dana jest wzorem [Anselin, Bao 1997, s. 35-59; Korol 2007, s. 92]:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\mathbf{z}^T\mathbf{W}\mathbf{z}}{\mathbf{z}^T\mathbf{z}}, \tag{2}$$

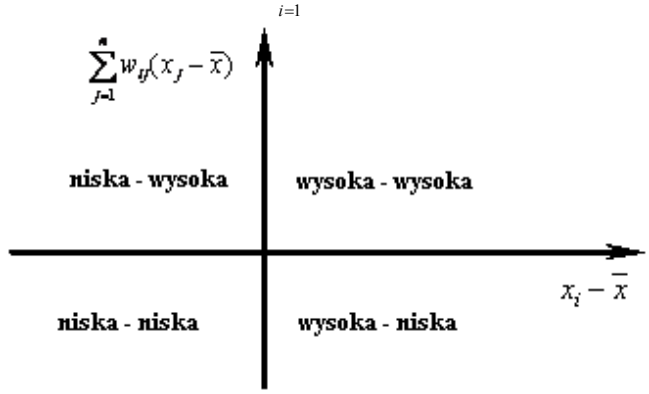
gdzie: n – liczba obserwacji; x_i, x_j – wartości zmiennej x w lokalizacjach i i j; $\bar{x}$ – średnia wartość obserwacji x_i; w_{ij} – elementy przestrzennej macierzy wag **W**[1]; **z** – wektor, który przyjmuje postać: $\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_n \end{bmatrix}$, gdzie $z_i = x_i - \bar{x}$.

[1] Przykładowa macierz sąsiedztwa jest standaryzowana wierszami i zawiera elementy w_{ij} = 1, gdy jednostki i oraz j mają wspólną granicę, 0 − w przeciwnym wypadku. Więcej o przestrzennych macierzach wag i odległości np. w: [Anselin 1988].

Globalna statystyka Morana I jest ważonym współczynnikiem korelacji podobnym do współczynnika korelacji Pearsona, służącym do wykrywania odchyleń w losowym rozkładzie zmiennej *x* w sensie przestrzennym [Suchecka 2002, s. 58]. Wykorzystywana jest do ustalenia, czy sąsiadujące ze sobą obszary są bardziej do siebie podobne (w sensie wartości zmiennej *x*) niż wynikałoby to ze stochastycznego charakteru badanego zjawiska. Macierz **W** opisuje stan rzeczywisty. Duża waga odpowiada dużej rzeczywistej korelacji, wtedy wartość statystyki Morana I jest wysoka. Jeżeli sąsiadujące obiekty przestrzenne (kraje, regiony, województwa, powiaty) są do siebie podobne, czyli tworzą klastry, wartość statystyki jest dodatnia. Jeżeli obiekty są różne (ich układ w przestrzeni jest regularny, nie tworzą skupień), to wartość statystyki jest ujemna. Gdy korelacja między sąsiadującymi wartościami nie występuje, wówczas wartość oczekiwana I opisana wzorem:

$$E(I) = \frac{-1}{(1-n)} \tag{3}$$

jest bliska 0 (podczas gdy liczba obserwacji *n* wzrasta). Wartość statystyki Morana I należy do przedziału [–1,1]. Zazwyczaj osiąga ona wartości $|I| < 1$. W celu ustalenia występowania zjawiska autokorelacji przestrzennej przeprowadza się testy randomizacji [Anselin 2003, s. 14][2]. W celu określenia typu autokorelacji przestrzennej dokonuje się analizy moranowskiego wykresu rozproszenia (*Moran scatterplot*). Wykres ten przedstawia liniowy związek pomiędzy *i*-tą wartością $(x_i - \bar{x})$ a średnią ważoną sąsiedztwa $\sum_{i=1}^{n} w_{ij}(x_i - \bar{x})$ (zob. rys. 1).



Rys. 1. Moranowski wykres rozproszenia
Źródło: opracowanie własne na podstawie [Anselin 2005, s. 129-133].

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

Zgrupowanie obserwacji w pierwszej lub trzeciej ćwiartce (*wysoka-wysoka*, *niska-niska*), wskazuje na występowanie autokorelacji dodatniej, natomiast w drugiej lub czwartej (*wysoka-niska*, *niska-wysoka*) − autokorelacji ujemnej.

Prócz konieczności badania globalnej autokorelacji przestrzennej w literaturze wskazuje się, że do uzyskania szczegółowego obrazu badanego zjawiska niezbędna jest analiza lokalnej autokorelacji przestrzennej LISA. Polega ona na badaniu korelacji wartości zmiennej w wybranej lokalizacji z jej sąsiadami [Anselin 1988, s. 284]. Lokalna statystyka Morana I_i wyrażona jest wzorem [Le Gallo, Ertur 2003, s. 175-201]:

$$I_i = \frac{(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{j=1}^{n} w_{ij}(x_j - \bar{x}), \tag{4}$$

gdzie: n – liczba obserwacji; x_i, x_j – wartości zmiennej x w lokalizacjach i i j; $\bar{x}$ – średnia wartość obserwacji x_i; w_{ij} – elementy przestrzennej macierzy wag **W**.

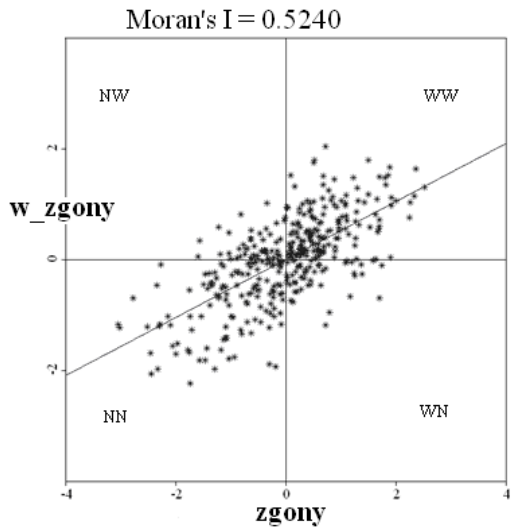
Obliczenie statystyki I_i dla wszystkich obserwacji umożliwia sporządzenie tzw. mapy istotności przedstawiającej wartość pseudowspółczynnika istotności testu randomizacyjnego wykonanego dla poszczególnych lokalizacji.

3. Zastosowanie ESDA w analizie umieralności z powodu chorób układu krążenia

Choroby układu krążenia są najczęstszą przyczyną zgonów w Polsce. Rocznie zawału doznaje około 80 tysięcy Polaków, 30% z nich umiera przed dotarciem do szpitala [*Encyklopedia zdrowia*… 2002, s. 1].

W opracowaniu analizie poddano wskaźnik wyrażony jako udział zgonów z powodu chorób układu krążenia w ogólnej liczbie zgonów w 2006 r. w Polsce. Badanie przeprowadzono dla 379 powiatów. Źródłem danych statystycznych był GUS. Celem badania była identyfikacja interakcji przestrzennych między liczbą zgonów w poszczególnych powiatach. Wykorzystano w nim narzędzia eksploracyjnej analizy danych przestrzennych opisane w drugiej części opracowania, a niezbędne obliczenia wykonano w pakiecie GeoDa.

Uzyskane wartości globalnych i lokalnych statystyk Morana pokazują przestrzenne zróżnicowanie poziomu śmiertelności z powodu chorób układu krążenia w przekroju powiatów w Polsce. Wartość globalnej statystyki Morana I, obliczona na podstawie macierzy sąsiedztwa (zbudowanej dla 5 najbliższych sąsiadów), wyniosła $I = 0{,}5240$, dla poziomu pseudo istotności $p = 0{,}001$ (test randomizacji dla 1000 losowych permutacji). Odrzucono hipotezę zerową mówiącą o braku autokorelacji przestrzennej, wnioskować należy zatem, że autokorelacja przestrzenna między badanymi jednostkami występuje i jest dodatnia. W przypadku umieralności z powodu chorób układu krążenia występuje więc tendencja do skupiania podobnych wartości wskaźnika w przestrzeni, co prezentuje moranowski wykres rozproszenia (zob. rys. 2).



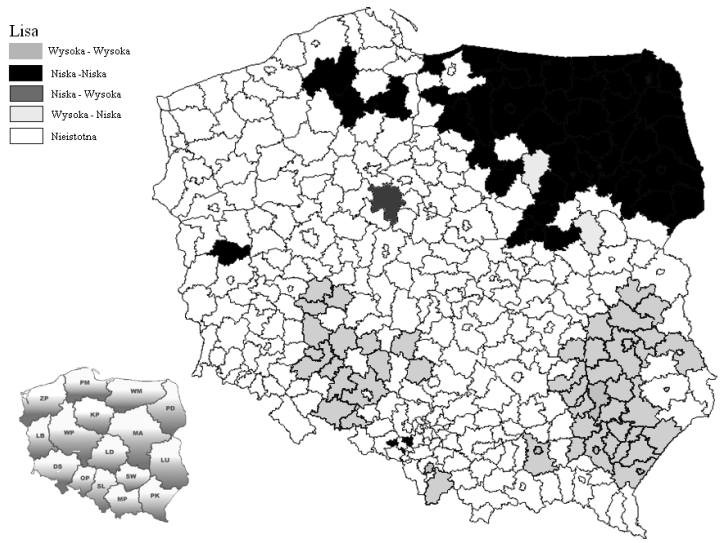
Rys. 2. Moranowski wykres rozproszenia dla analizowanej zmiennej w przekroju powiatów w 2006 r. w Polsce[3]
Źródło: opracowanie własne w pakiecie GeoDa.

Występowanie przestrzennej autokorelacji dodatniej oznacza, że istnieją pewne zależności przestrzenne w kształtowaniu się poziomu śmiertelności z powodu chorób układu krążenia w Polsce. Można wnioskować, że spadek (wzrost) liczby zgonów w jednym powiecie może wpływać na spadek (wzrost) śmiertelności w powiatach sąsiednich. W Polsce od roku 2001 następuje spadek umieralności z powodu zawałów i innych chorób układu krążenia. Tendencja ta jest sukcesem polskich kardiologów oraz efektem pomocy państwa w dostępności do usług kardiologicznych. Poprawa nastąpiła zwłaszcza dzięki realizowanemu Narodowemu Programowi Ochrony Serca (lata 1993-2001). W tym czasie szpitale zakupiły nowoczesną aparaturę, a lekarze mogli się szkolić za granicą. Zatem większy dostęp do specjalistycznej aparatury i lekarzy w jednym powiecie najprawdopodobniej powoduje migracje ludności z powiatów sąsiednich, a tym samym zmniejszenie liczby zgonów wynikających z chorób układu krążenia. W kolejnym etapie badania wyznaczono lokalne wartości statystyki Morana I_i (LISA). Na podstawie uzyskanych wyników można wyciągnąć wnioski dotyczące korelacji przestrzennej poszczególnych powiatów z ich sąsiadami (zob. rys. 3).

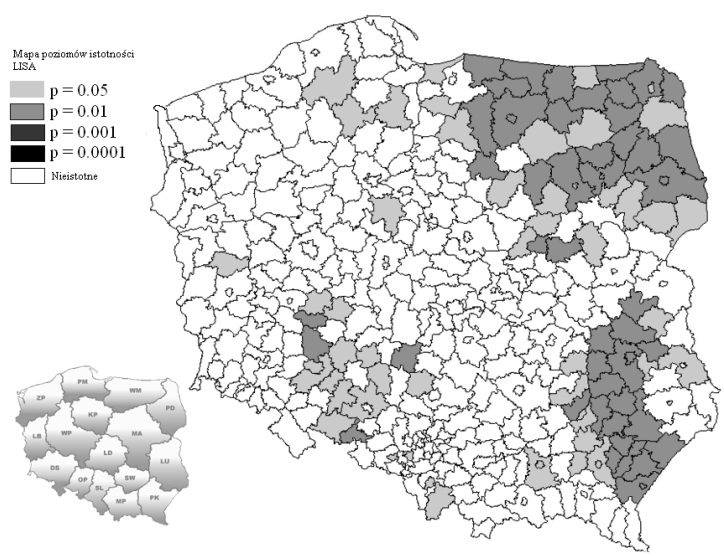
Rysunek 4 przedstawia mapę poziomów istotności (p) dla poszczególnych wartości statystyki LISA, które zostały policzone dla 1000 permutacji w teście randomizacji. Najciemniejszy odcień odpowiada najniższej wartość współczynnika pseudo-

[3] Zmienna „zgony" oznacza udział zgonów z powodu chorób układu krążenia w ogólnej liczbie zgonów w 2006 r. w Polsce, zmienna „w_zgony" to opóźniona przestrzennie zmienna „zgony" (wartości zmiennej „zgony" zostały przemnożone przez przestrzenną macierz wag **W**).

istotności, co oznacza wysoką korelację z sąsiadami. Najbardziej skorelowane ze sobą są powiaty z północno-wschodniej i południowo-wschodniej części Polski.



Rys. 3. Typy zależności przestrzennych na podstawie wartości LISA − umieralność z powodu chorób układu krążenia w powiatach w 2006 r.
Źródło: opracowanie własne w pakiecie GeoDa.



Rys. 4. Mapa istotności dla wyznaczonych wartości statystyki LISA
Źródło: opracowanie własne w pakiecie GeoDa.

Wysoka wartość globalnej statystyki Morana I znajduje potwierdzenie w obrazie uzyskanym na podstawie LISA (zob. rys. 3 i 4). Na rysunku 3 wyraźnie widać skupiska przestrzenne (klastry) obiektów o podobnych wartościach zmiennej. Wysokim udziałem zgonów z powodu chorób układu krążenia charakteryzują się powiaty znajdujące się w województwach lubelskim, podkarpackim, opolskim i dolnośląskim oraz miasto Zamość. Regiony te sąsiadują z obszarami o równie wysokim poziomie badanego zjawiska, jednocześnie przestrzennie ze sobą korelują.

W Polsce jest zbyt mało oddziałów kardiologicznych. Według lekarzy główne problemy to długie oczekiwanie na dostęp do usług specjalistycznych oraz dysproporcje w rozmieszczeniu ośrodków intensywnej opieki kardiologicznej – w trzech województwach nie ma ich w ogóle, są to: lubelskie, podkarpackie i dolnośląskie [Broda 2004, s. 1-3].

W północno-wschodniej części kraju (w woj. warmińsko-mazurskim, pomorskim i podlaskim) występują obszary o niskich, homogenicznych wartościach badanej zmiennej. Powiaty w tych województwach charakteryzują się niskim poziomem umieralności z powodu chorób układu krążenia. Niskie wartości badanego zjawiska występują również w powiatach: rybnickim, jastrzębskim, świętochłowicki (stanowią one odosobnione obszary; zob. rys. 3). Zróżnicowanie przestrzenne w zakresie analizowanej zmiennej może być wynikiem występowania w niektórych regionach (w woj. śląskim i mazowieckim) całodobowych oddziałów intensywnego leczenia ostrych zespołów wieńcowych.

Warto również zwrócić uwagę na pojawiające się jednostki odstające. Powiatem o niskiej wartości wskaźnika umieralności z powodu chorób układu krążenia jest powiat inowrocławski. Sąsiaduje on z jednostkami o wysokich wartościach badanego zjawiska. Natomiast powiaty wągrowiecki i mławski charakteryzują się wysokimi wartościami wskaźnika, a otoczone są jednostkami o niskim poziomie śmiertelności (zob. rys. 3).

4. Podsumowanie

Obecnie jesteśmy zasypywani danymi statystycznymi. Niemal każda instytucja zbiera dane. Niestety większość z nich marnuje się z powodu zbyt małej liczby specjalistów − analityków, którzy potrafią wydobyć z nich wiedzę. Efektywna analiza danych przestrzennych pozwala sformułować cenne wnioski mogące być niezwykle istotne przy podejmowaniu decyzji w zakresie gospodarki przestrzennej na szczeblu zarówno lokalnym, jak i państwowym.

Z analizy przeprowadzonej w opracowaniu wynika, że w Polsce w roku 2006 występowało znaczne zróżnicowanie umieralności z powodu chorób układu krążenia między województwami i powiatami. Można wysnuć wniosek, że województwa lubelskie, podkarpackie, opolskie i dolnośląskie wymagają restrukturyzacji i modernizacji służby zdrowia w zakresie leczenia chorób układu krążenia. Wyniki analizy pokazały również, że wystarczy stworzenie w jednym powiecie profesjo-

nalnego ośrodka kardiologicznego, aby w całym sąsiednim obszarze spadła liczba zgonów z powodu chorób układu krążenia.

Literatura

Anselin L. (2005), *Exploring spatial data with GeoDATM: a workbook*, Spatial Analysis Laboratory, http://sal.uiuc.edu/.

Anselin L. (2003), *GeoDa 0.9 user's guide*, Spatial Analysis Labolatory, http://sal.uiuc.edu.

Anselin L. (1988), *Spatial econometrics: methods and models*, Kluwer Academic Publishers, Dordrecht.

Anselin L., Bao S. (1997), *Exploratory spatial data analysis*, [w:] *Recent developments in spatial analysis*, Springer-Verlag, Berlin.

Broda G. (2004), *Choroby układu krążenia w Polsce – umieralność*, Instytut Kardiologii w Warszawie.

Encyklopedia zdrowia (2002), *Kardiolodzy: spadła umieralność z powodu chorób układu krążenia*, www.zdrowie.med.pl.

Korol J. (2007), *Wskaźniki zrównoważonego rozwoju w modelowaniu procesów regionalnych*, Wydawnictwo Adam Marszałek, Toruń.

Larose D.T. (2006), *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, PWN, Warszawa.

Le Gallo J., Ertur C. (2003), *Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980-1995*, Papers in Regional Science v. 82 (2).

Suchecka J. (2002), *Metody statystyczne. Zarys teorii i zadania*, Politechnika Częstochowska, Częstochowa.

Tobler W.R. (1970), *A computer movie simulating urban growth in the Detroit region*, "Economic Geography" 46, s. 234-240.

THE APPLICATION OF EXPLORATORY SPATIAL DATA ANALYSIS IN RESEARCH ON THE MORTALITY LEVEL IN POLAND

Summary

The aim of this paper is to present selected exploratory spatial data analysis methods (ESDA). ESDA is a series of statistical techniques used in a spatial dependence analysis. First, we described basic methods of local and global statistics of spatial autocorrelation. Subsequently, we presented examples of ESDA applications in analysis of mortality caused by circulatory system diseases. This analysis was done at the poviat level in Poland in 2006.