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

Jacek Batóg
Uniwersytet Szczeciński

WYKORZYSTANIE ANALIZY DYSKRYMINACYJNEJ Z AUTOKORELACJĄ PRZESTRZENNĄ DO KLASYFIKACJI OBIEKTÓW

1. Wstęp

Jedną z powszechnie stosowanych procedur klasyfikacji obiektów społeczno-gospodarczych jest analiza dyskryminacyjna. Ponieważ podejście to doczekało się wielu opracowań o charakterze zarówno metodologicznym, jak i empirycznym, pominiemy opis samej metody, a także charakterystykę jej założeń[1].

W pracy podjęto próbę modyfikacji klasycznego podejścia stosowanego w analizie dyskryminacyjnej przez wprowadzenie do niej czynnika przestrzennego. Czynnik ten ma odzwierciedlać występowanie zjawiska autokorelacji przestrzennej między badanymi obiektami w przestrzeni społeczno-gospodarczej. Pozwoli to na uwzględnienie powiązań o charakterze przestrzennym między obiektami w procesie ich przyporządkowywania do danej populacji (grupy). Podstawowym celem tej modyfikacji, spotykanej również w przypadku innych procedur statystycznych i ekonometrycznych[2], jest ocena, czy wprowadzenie efektu autokorelacji przestrzennej do klasycznej procedury klasyfikacyjnej doprowadzi do wzrostu liczby poprawnie zakwalifikowanych przypadków.

Efektywność zaproponowanych rozwiązań zostanie zweryfikowana na podstawie rezultatów badania polegającego na klasyfikacji gmin dwóch województw: zachodniopomorskiego i kujawsko-pomorskiego.

2. Autokorelacja przestrzenna w modelowaniu procesów ekonomicznych

Zjawisko występowania istotnego podobieństwa zmiennych charakteryzujących sąsiadujące obiekty może być spowodowane nie tylko efektami o charakterze ekono-

[1] Podstawowe informacje o analizie dyskryminacyjnej znaleźć można między innymi w pracach: [Jajuga 1990; Krzyśko 1990; Wawrzyniak, Batóg 1997].

[2] Jako przykład można tu podać modele przestrzenne stosowane w analizie regresji [Anselin 1988].

micznym, demograficznym, technologicznym, klimatycznym i instytucjonalnym, lecz również czynnikami historycznymi, kulturowymi i socjologicznymi [Zeliaś 1991, s. 97]. Związki istniejące między obiektami w przestrzeni mogą istotnie wpływać na wyniki przeprowadzanych analiz ilościowych. Jak pisze A. Kopczewska: „Dzięki estymacji modeli uwzględniających czynnik przestrzenny możliwe jest określenie przestrzennej zależności pomiędzy obserwacjami w różnych lokalizacjach, a także udowodnienie, że istnieje niemierzalny czynnik przestrzenny różnicujący badane zjawisko pomiędzy lokalizacjami" [Kopczewska 2006, s. 13]. Przestrzeń wymieniana jest jako jedna z pięciu sił sprawczych kształtujących zdarzenia ekonomiczne [Hozer 2003].

Występowanie autokorelacji przestrzennej może być również związane z konstrukcją stosowanych procedur. Najczęściej wskazuje się w tym przypadku na występowanie błędów specyfikacji, czyli pominięcie istotnej zmiennej objaśniającej charakteryzującej się autokorelacją przestrzenną [Fischer, Stirbrock 2006, s. 701], oraz na pojawianie się błędów pomiaru wynikających z niezgodności podziałów administracyjnych obiektów z rozmieszczeniem procesów generujących dane statystyczne [LeSage 1999, s. 3].

3. „Przestrzenna" analiza dyskryminacyjna

Wielu autorów zajmujących się zagadnieniami klasyfikacyjnymi zauważa konieczność uwzględnienia w procedurach analizy wielowymiarowej powiązań istniejących między obiektami sąsiadującymi w danej przestrzeni oraz wyraża opinię, że tego typu modyfikacje pozwalają uzyskiwać bardziej adekwatne wyniki klasyfikacji[3]. Proponowane przez nich rozwiązania różnią się przede wszystkim sposobem adaptacji czynnika przestrzennego. Najczęściej stosowane są w tym zakresie dwa alternatywne podejścia: dokonywanie korekty prawdopodobieństw *a priori* [Cutillo, Amato 2008] oraz modyfikacja prawdopodobieństw *a posteriori* [Steele, Redmond 2001]. W tym drugim przypadku korekcie podlega formuła Bayesa poprzez wprowadzenie wag przestrzennych określających przynależność poszczególnych obiektów do danych grup na podstawie ich lokalizacji w rozpatrywanej przestrzeni.

Proponowana obecnie procedura uwzględniania zależności przestrzennych w funkcji dyskryminacyjnej w zagadnieniach klasyfikacyjnych ma nieco odmienny charakter. Polega ona na korekcie macierzy wartości zmiennych diagnostycznych w sposób, u podstaw którego leży założenie, że oprócz oryginalnych wartości zmiennych charakteryzujących badane obiekty o przynależności określonego obiektu do danej grupy decydują również charakterystyki obiektów sąsiednich. Ten drugi czynnik dla danego obiektu konstruowany jest jako średnia wartość zmiennych diagnostycznych charakteryzujących obiekty, które mają z nim wspólne granice. Jak widać, dokonujemy w ten sposób transformacji postaci funkcji dyskryminacyjnej przez wprowadzenie dodatkowych zmiennych diagnostycznych uwzględniających powiązania przestrzenne.

[3] Zob. na przykład [Thioulouse, Chessel, Champely 1995, s. 2; Dučinskas, Šaltytė 2002].

Ocenę zasadności przedstawionego rozwiązania przeprowadzamy, porównując trafność klasyfikacji uzyskaną w przypadku zastosowania klasycznej funkcji dyskryminacyjnej:

$$Y_i = \alpha_0 + \sum_{j=1}^{k} X_{ij} \tag{1}$$

z modelem o postaci:

$$Y_i = \alpha_0 + \sum_{j=1}^{k} WX_{ij} + \sum_{j=1}^{k} X_{ij}\,, \tag{2}$$

gdzie: W – macierz wag przestrzennych.

Macierz wag przestrzennych W wykorzystywana w procedurze estymacji parametrów „przestrzennych" funkcji dyskryminacyjnych ma zera na głównej przekątnej, a poza główną przekątną − elementy określające liczbę powiązań badanych obiektów z obiektami sąsiadującymi. Poszczególne zawarte w wierszach elementy tej macierzy poddawane są standaryzacji zapewniającej spełnienie warunku[4]:

$$\sum_{j=1}^{k} d_{ij} = 1\,. \tag{3}$$

4. Przykład empiryczny

Ocena zaproponowanego podejścia przeprowadzona zostanie przez porównanie wyników klasyfikacji gmin województwa zachodniopomorskiego i województwa kujawsko-pomorskiego uzyskanych z wykorzystaniem klasycznej analizy dyskryminacyjnej oraz jej wariantu przestrzennego do trzech grup obiektów[5]: gmin miejskich (M), gmin miejsko-wiejskich (MW) oraz gmin wiejskich (W). W zbiorze cech diagnostycznych występuje 13 zmiennych charakteryzujących wybrane aspekty funkcjonowania tych gmin w roku 2006. Należą do nich: dochody własne *per capita* (X_1), wydatki inwestycyjne *per capita* (X_2), lesistość (X_3), ludność w wieku nieprodukcyjnym na 100 osób w wieku produkcyjnym (X_4), przyrost naturalny na 1000 ludności (X_5), gęstość zaludnienia (X_6), powierzchnia nowo oddanych mieszkań na 1000 ludności (X_7), osoby fizyczne prowadzące działalność gospodarczą na 1000 ludności (X_8), podmioty zarejestrowane w systemie REGON na 1000 ludności (X_9), pracujący na 1000 ludności (X_{10}), pracujący w usługach na 1000 ludności (X_{11}), nakłady na ochronę środowiska i gospodarkę wodną na 1000 ludności (X_{12}) oraz zużycie wody z wodociągów w gospodarstwach domowych na 1 mieszkańca (X_{13}).

[4] Zob. [Abreu, de Groot, Florax 2004].

[5] Dyskryminacji nie podlegały miasta na prawach powiatu. Obiekty te były jednak brane pod uwagę przy wyznaczaniu macierzy wag przestrzennych.

W procesie budowy funkcji dyskryminacyjnych weryfikacji poddano standardowe założenia analizy dyskryminacyjnej. Mimo że nie wszystkie te założenia zostały spełnione, to jak wskazuje wielu autorów, nie musi to wpływać negatywnie na efekty procesu dyskryminacji [Domański, Misztal 1998; Gatnar 1998].

Poniżej dla obu rozpatrywanych województw przedstawiono kanoniczne funkcje dyskryminacyjne uzyskane w wariancie klasycznym[6], podstawowe miary charakteryzujące te funkcje (tab. 1) oraz macierze klasyfikacji (tab. 2):

1. Województwo zachodniopomorskie:

– funkcja 1 (X_{10}, X_6, X_1, X_7 i X_{11}):

$$\hat{y}_i = -20,71 - 0,0004X_1 + 0,00004X_2 - 0,0166X_3 - 0,0102X_4 - 0,0354X_5 + 0,1926X_6 - 0,0007X_7 - 0,0022X_9 + 0,0180X_{10} + 1,9166X_{11} - 0,00001X_{12} - 0,0011X_{13},$$

– funkcja 2 (X1, X9, X2, X3 i X7):

$$\hat{y}_i = 6,61 \cdot 0,0011X_1 + 0,00114X_2 - 0,0546X_3 + 0,0838X_4 + 0,0638X_5 - 0,0898X_6 + 0,0015X_7 + 0,0186X_9 + 0,0025X_{10} + 0,0092X_{11} - 0,00008X_{12} - 0,0713X_{13}.$$

2. Województwo kujawsko-pomorskie:

– funkcja 1 (X9, X8, X6 i X1):

$$\hat{y}_i = 6,81 + 0,0015X_1 + 0,0004X_2 + 0,0137X_3 - 0,0510X_4 + 0,0342X_5 - 0,0050X_6 - 0,0004X_7 + 0,1548X_8 - 0,1833X_9 - 0,0027X_{10} + 0,2632X_{11} - 0,0005X_{12} - 0,0091X_{13},$$

– funkcja 2 (X10, X8, X9 i X11):

$$\hat{y}_i = 3,82 - 0,00002X_1 - 0,0009X_2 - 0,0128X_3 - 0,0808X_4 - 0,0682X_5 - 0,0028X_6 + 0,0011X_7 + 0,00440X_8 - 0,0317X_9 + 0,0187X_{10} + 3,1949X_{11} + 0,0003X_{12} - 0,0538X_{13}.$$

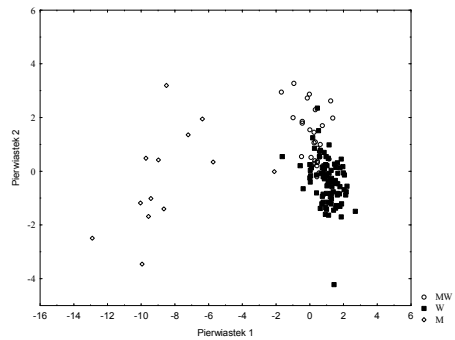
Dla obu województw pierwsza funkcja dyskryminuje gminy miejskie od pozostałych, a druga odpowiedzialna jest za odróżnianie gmin miejsko-wiejskich od wiejskich (zob. rys. 1). Na podstawie standaryzowanych parametrów określono zmienne o znacznym wpływie na wartości funkcji dyskryminacyjnych – zostały one zaznaczone w nawiasach.

Tabela 1. Podsumowanie rezultatów estymacji klasycznych funkcji dyskryminacyjnych

Funkcja	R	λ Wilksa	χ^2	Stopnie swobody	p	Wartości własne	Wariancja (%)
Zachodniopomorskie							
1	0,77	0,31	118,95	24	0,0000	1,44	0,82
2	0,48	0,76	27,52	11	0,0038	0,31	0,18
Kujawsko-pomorskie							
1	0,94	0,08	328,21	26	0,0000	7,49	0,94
2	0,55	0,69	47,97	12	0,0000	0,44	0,06

Źródło: obliczenia własne.

[6] W równaniach podane zostały surowe współczynniki tych funkcji.

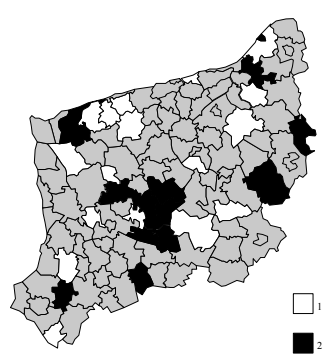


Rys. 1. Rozrzut punktów obrazujących gminy województwa kujawsko-pomorskiego uzyskany na podstawie wartości klasycznych funkcji dyskryminacyjnych
Źródło: opracowanie własne.

Tabela 2. Macierze klasyfikacji (wariant klasyczny)

Zachodniopomorskie				
	trafność klasyfikacji (%)	MW	W	M
MW	66,67	34	15	2
W	73,08	14	38	0
M	87,50	1	0	7
Razem	71,17	49	53	9
Kujawsko-pomorskie				
	trafność klasyfikacji (%)	MW	W	M
MW	48,57	17	18	0
W	94,57	5	87	0
M	92,31	1	0	12
Razem	82,85	23	105	12

Źródło: obliczenia własne.



Rys. 2. Niepoprawnie sklasyfikowane gminy w województwie zachodniopomorskim w wariancie klasycznym (1 – klasyfikacja „zawyżona", 2 – klasyfikacja „zaniżona")
Źródło: opracowanie własne.

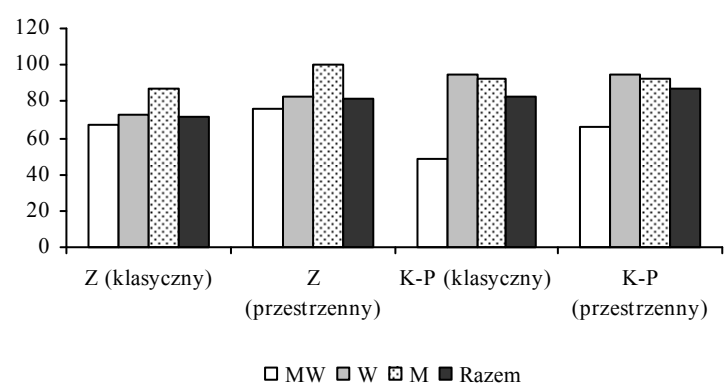
Zaprezentowane wyniki świadczą o zdecydowanej dominacji pierwszych funkcji dyskryminacyjnych w zakresie wielkości wyjaśnianej wariancji. Widoczne jest również poprawniejsze klasyfikowanie gmin w przypadku województwa kujawsko-pomorskiego. Uwagę jednak w tym przypadku zwraca stosunkowo duża liczba błędnie sklasyfikowanych gmin miejsko-wiejskich, z których ponad połowa została przypisana do grupy gmin wiejskich.

W tabeli 3 zamieszczono macierze klasyfikacji uzyskane w wariancie przestrzennym. Porównanie poprawnie sklasyfikowanych przypadków w obu rozpatrywanych wariantach pozwala stwierdzić, że uwzględnienie w funkcji dyskryminacyjnej powiązań przestrzennych znacznieo podwyższa liczbę poprawnie sklasyfikowanych obiektów (zob. rys. 3).

Tabela 3. Macierze klasyfikacji (wariant przestrzenny)

Zachodniopomorskie				
	trafność klasyfikacji (%)	MW	W	M
MW	76,47	39	12	0
W	82,69	9	43	0
M	100,00	0	0	8
Razem	81,08	48	55	8
Kujawsko-pomorskie				
	trafność klasyfikacji (%)	MW	W	M
MW	65,71	23	12	0
W	94,57	5	87	0
M	92,31	1	0	12
Razem	87,14	29	99	12

Źródło: obliczenia własne.



Rys. 3. Porównanie trafności klasyfikacji gmin (%)
Źródło: opracowanie własne.

Poprawa ta dotyczy wszystkich kategorii gmin województwa zachodniopomorskiego oraz gmin miejsko-wiejskich województwa kujawsko-pomorskiego.

5. Wnioski

Otrzymane rezultaty wskazują na zasadność uwzględniania w zagadnieniach klasyfikacji powiązań przestrzennych między badanymi obiektami. Pozwala to na redukcję popełnianych błędów, które w analizowanym przypadku są spowodowane przede wszystkim niezgodnością sytuacji społeczno-gospodarczej niektórych gmin z przypisanym im statusem administracyjnym. Sytuacje tego typu mogą wynikać między innymi z faktu, że gminy wiejskie oraz miejsko-wiejskie sąsiadujące z gminami miejskimi mogą się charakteryzować zbliżonym poziomem zmiennych diagnostycznych do tych ostatnich. W celu pełniejszej weryfikacji skuteczności proponowanej procedury niezbędne jest jednak przeprowadzenie szerszych analiz tego typu oraz dokonanie oceny, jak zmienia się trafność klasyfikacji wraz ze zmianą poziomu autokorelacji przestrzennej cech diagnostycznych.

Literatura

Abreu M., de Groot H.L.F., Florax R.J.G.M. (2004), *Space and growth: a survey of empirical evidence and methods*, Tinbergen Institute Discussion Paper no TI 2004-129/3, Amsterdam.

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

Cutillo L., Amato U. (2008), *Localized empirical discriminant analysis*, "Computational Statistics and Data Analysis", vol. 52, issue 11, s. 4966-4978.

Domański C., Misztal M. (1998), *Zastosowanie wybranych metod dyskryminacji do wspomagania diagnozy i określania ryzyka operacyjnego u pacjentów z chorobą wieńcową*, [w:] *Modelowanie preferencji a ryzyko '98*, red. T. Trzaskalik, AE, Katowice.

Dučinskas K., Šaltytė J. (2002), *The effect of spatial autocorrelation on the error rates of the linear discriminant functions*, "Lithuanian Mathematical Journal", vol. 42, no 2, s. 133-139.

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

Fischer M.M., Stirböck C. (2006), *Pan-european regional income growth and club-convergence. Insights from a spatial econometric perspective*, "Annals of Regional Science", vol. 40, s. 693-721.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.

Hozer J. (2003), *Tempus locus homo casus et fortuna regit factum. Zbiór esejów ekonomicznych*, Instytut Analiz, Diagnoz i Prognoz Gospodarczych w Szczecinie, Oficyna „in Plus", Szczecin.

Jajuga K. (1990), *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa.

Kopczewska K. (2006), *Ekonometria i statystyka przestrzenna z wykorzystaniem programu R CRAN*, CeDeWu Sp. z o.o., Warszawa.

Krzyśko M. (1990), *Analiza dyskryminacyjna*, Wydawnictwo Naukowo-Techniczne, Warszawa.

LeSage J.P. (1999), *Spatial econometrics*, University of Toledo, Toledo.

Steele B.M., Redmond R.L. (2001), *A method of exploiting spatial information for improving classification rules: application to the construction of polygon-based land cover maps*, "International Journal of Remote Sensing", vol. 22, no 16, s. 3143-3166.

Thioulouse J., Chessel D., Champely S. (1995), *Multivariate analysis of spatial patterns: a unified approach to local and global structures*, "Environmental and Ecological Statistics", vol. 2, no 1, s. 1-14.

Wawrzyniak K., Batóg B. (1997), *Wykorzystanie funkcji dyskryminacyjnej do oceny kondycji finansowo-ekonomicznej spółek i przedsiębiorstw I, II, III i IV transzy, alokowanych do Narodowych Funduszy Inwestycyjnych*, „Przegląd Statystyczny" nr 1.

THE APPLICATION OF THE DISCRIMINANT ANALYSIS WITH THE SPATIAL AUTOCORRELATION TO THE CLASSIFICATION OF OBJECTS

Summary

The aim of the paper is the presentation of the modification of the classical discriminant analysis. This proposal is connected with the occurrence of the spatial autocorrelation. The spatial autocorrelation is described by matrix of spatial weights. Its elements represent directly nonmeasurable spatial relations between discriminated objects. This approach is an alternative solution to existing methods based on correcting of *a priori* or *a posteriori* probabilities. The empirical verification of the proposed method testifies the significant improvement of the accuracy of the classification when the spatial discriminant analysis is applied.