PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 469 • 2017

Taksonomia 29 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Elżbieta Zalewska
Uniwersytet Łódzki
e-mail: zalewska.e@uni.lodz.pl

ZASTOSOWANIE ANALIZY SKUPIEŃ I METODY PORZĄDKOWANIA LINIOWEGO W OCENIE POLSKIEGO SZKOLNICTWA WYŻSZEGO

APPLICATION OF CLUSTER ANALYSIS AND LINEAR ORDERING IN THE ASSESSMENT OF POLISH HIGHER EDUCATION

DOI: 10.15611/pn.2017.469.24
JEL Classification JEL: C38, A22

Streszczenie: W polskim szkolnictwie wyższym następuje wiele zmian, głównie wynikających z przemian demograficznych oraz wprowadzania do życia akademickiego nowoczesnych technologii. Szybki wzrost liczby uczelni i spadek liczby studentów powoduje konkurencyjność na polskich uczelniach wyższych. Celem pracy jest prezentacja wyników wielowymiarowej analizy porównawczej stanu polskiego szkolnictwa wyższego w 2014 roku w podziale na województwa na podstawie danych Głównego Urzędu Statystycznego i Banku Danych Lokalnych. Ocena obejmuje zarówno uczelnie publiczne, jak i niepubliczne oraz studia stacjonarne i niestacjonarne. Za pomocą analizy skupień pogrupowano województwa, uwzględniając podobieństwo stanu szkolnictwa wyższego, oraz zastosowano porządkowanie liniowe pozwalające ustalić klasyfikację województw. Do obliczeń wykorzystano pakiet Statistica.

Słowa kluczowe: szkolnictwo wyższe, analiza skupień, porządkowanie liniowe.

Summary: Polish academic education faces significant changes, mostly resulting from introducing technologically advanced ways of teaching and demographic change. A rapid increase of the number of universities and a decrease of the number of students cause competitiveness among Polish universities. The aim of the article is the presentation of the results of multidimensional comparative analysis of conditions of Polish higher education in 2014 per voivodeships, based on data from the Central Statistical Office and the Local Data Bank. For grouping voivodeships, cluster analysis was used, taking into account the similarity of the condition of higher education and sorting line for determining the classification of voivodeships. Calculations were prepared using Statistica software.

Keywords: higher education, cluster analysis, linear ordering.

1. Wstęp

W XXI w. zmieniają się oczekiwania i wymagania wobec uczelni wyższych. Zmiany demograficzne, intensywny rozwój technologii, globalizacja powodują konkurencyjność na polskich uczelniach.

Analizując polskie szkolnictwo wyższe, należy pamiętać o wewnętrznym zróżnicowaniu, ocena obejmuje zarówno uczelnie publiczne, jak i niepubliczne oraz studia stacjonarne i niestacjonarne. W roku akademickim 2014/2015 w Polsce funkcjonowały 434 uczelnie wyższe, w tym 302 to uczelnie niepubliczne. Wszystkie uczelnie kształciły ponad 1 470 000 studentów, w tym około 24% studentów przypadało na uczelnie niepubliczne.

Celem pracy jest pogrupowanie województw w skupienia o podobnym stanie szkolnictwa wyższego oraz utworzenie rankingu najefektywniejszych województw ze względu na wybrane cechy.

2. Charakterystyka materiału badawczego

Ze względu na dostępność danych, analiza statystyczna wybranych zmiennych objęła rok 2014. Materiał empiryczny pochodzi z Banku Danych Lokalnych (BDL), publikacji Głównego Urzędu Statystycznego (GUS) *Szkoły wyższe i ich finanse w 2014 r.* oraz statystyk konkursów NCN. Rozwój szkolnictwa wyższego w Polsce scharakteryzowano za pomocą następujących zmiennych:

X_1 – liczba absolwentów szkół wyższych na 10 tys. ludności,

X_2 – liczba profesorów na wszystkich nauczycieli akademickich,

X_3 – słuchacze studiów podyplomowych na 10 tys. ludności,

X_4 – uczestnicy studiów doktoranckich na 10 tys. ludności,

X_5 – odsetek studiujących na kierunkach technicznych i przyrodniczych, bez cudzoziemców,

X_6 – liczba szkół wyższych na 10 tys. ludności w wieku 20-24 lata,

X_7 – udział bezrobotnych zarejestrowanych z wyższym wykształceniem w liczbie ludności w wieku produkcyjnym,

X_8 – liczba studentów otrzymująca stypendia rektora dla najlepszych studentów na 10 tys. studentów ogółem (z cudzoziemcami), stan w dn. 30.11.2014,

X_9 – liczba studentów przypadająca na jednego nauczyciela akademickiego,

X_{10} – wysokości finansowania projektów NCN w przeliczeniu na liczbę zakwalifikowanych wniosków NCN [http://ncn.gov.pl/],

X_{11} – odsetek studentów cudzoziemców studiujących na polskich uczelniach.

W celu oceny wybranych zmiennych wyznaczono macierz korelacji pomiędzy zmiennymi. Zbyt silna zależność korelacyjna pomiędzy cechami może świadczyć o powielaniu przez te zmienne informacji. Zatem za progowy poziom współczynnika korelacji przyjęto ($r^* = 0{,}9$) [Strzała, Przechlewski 1994]. Najsilniejsza zależność

liniowa występuje pomiędzy liczbą uczestników studiów doktoranckich na 10 tys. ludności oraz liczbą absolwentów szkół wyższych na 10 tys. ludności ($r^* = 0{,}91$).

Tabela 1. Macierz korelacji badanych zmiennych

Zmienna	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
X_1	1,00	–0,49	0,64	**0,91**	0,21	0,49	0,05	–0,14	0,01	0,58	0,54
X_2	–0,49	1,00	0,11	–0,44	–0,37	–0,01	–0,15	–0,14	0,42	–0,38	–0,34
X_3	0,64	0,11	1,00	0,71	–0,21	0,77	0,01	–0,37	0,03	0,43	0,46
X_4	**0,91**	–0,44	0,71	1,00	0,16	0,57	–0,06	–0,32	–0,23	0,67	0,57
X_5	0,21	–0,37	–0,21	0,16	1,00	–0,04	0,12	–0,06	0,24	0,02	–0,03
X_6	0,49	–0,01	0,77	0,57	–0,04	1,00	0,01	–0,58	0,06	0,38	0,21
X_7	0,05	–0,15	0,01	–0,06	0,12	0,01	1,00	0,19	0,27	–0,34	0,46
X_8	–0,14	–0,14	–0,37	–0,32	–0,06	–0,58	0,19	1,00	–0,02	–0,34	0,06
X_9	0,01	0,42	0,03	–0,23	0,24	0,06	0,27	–0,02	1,00	–0,41	–0,09
X_{10}	0,58	–0,38	0,43	0,67	0,02	0,38	–0,34	–0,34	–0,41	1,00	0,43
X_{11}	0,54	–0,34	0,46	0,57	–0,03	0,21	0,46	0,06	–0,09	0,43	1,00

Źródło: opracowanie własne na podstawie BDL i GUS.

W wyniku eliminacji zmiennych silnie skorelowanych wybrano dziesięć zmiennych diagnostycznych: $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}$.

3. Analiza skupień

Analiza skupień to zbiór metod wielowymiarowej analizy statystycznej, polegająca na segmentacji danych w celu wyodrębnienia jednorodnych obiektów badanej populacji. Metoda ta polega na dzieleniu zbioru danych na grupy, tak aby uzyskać skupienia, w których elementy są do siebie podobne, a jednocześnie różne od elementów z pozostałych grup [Gatnar, Walesiak 2004].

Algorytmy analizy skupień możemy podzielić na metody hierarchiczne, niehierarchiczne oraz metody rozmytej analizy skupień. W metodach hierarchicznych budowa skupień polega na łączeniu elementów w podgrupy, tak by utworzyć dla obiektów porządek klasyfikacji, zaczynając od podziału, w którym każdy element stanowi odrębne skupienie, a kończąc na podziale, w którym wszystkie obiekty należą do jednego skupienia. Kolejna grupa metod to metody niehierarchiczne, do których należy metoda *k*-średnich, polegająca na podzieleniu zbioru na z góry założoną liczbę klas. Uzyskany podział jest analizowany i poprawiany, tak aby uzyskać minimalizację wariancji wewnątrz grup. Metody rozmytej analizy skupień, np. *c*-średnich, pozwalają przydzielić elementy do więcej niż jednej klasy z uwzględnieniem prawdopodobieństwa przynależności.

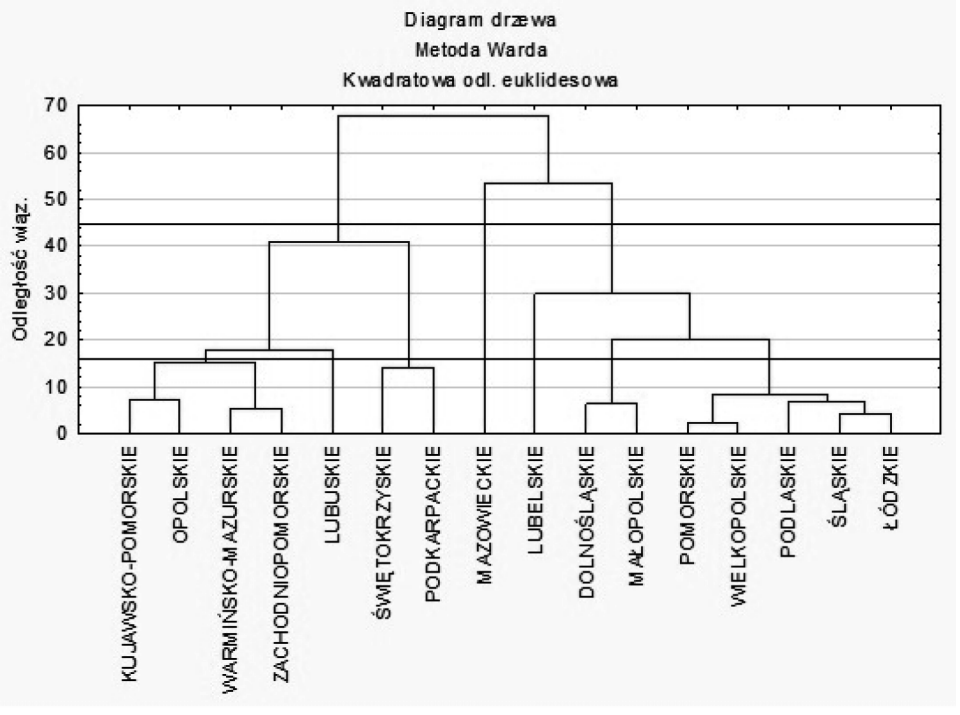
W przeprowadzonym badaniu wykorzystano technikę aglomeracyjną, będącą jedną z metod hierarchicznych. Odległość pomiędzy obiektami ustalono na podstawie kwadratu odległości euklidesowej według następującej formuły:

$$d(x,y) = \sum_{i=1}^{p} (x_i - y_i)^2. \tag{1}$$

W celu skorzystania z wyżej opisanej metryki przeprowadzono standaryzację zmiennych według poniższego wzoru

$$u_{ij} = \frac{x_{ij} - \bar{x}_j}{S_{x_j}},$$

gdzie: x_{ij} –wartość empiryczna *j*-tej zmiennej w *i*-tym województwie, $\bar{x}_j$ średnia arytmetyczna *j*-tej zmiennej, S_{xj} odchylenie standardowe *j*-tej zmiennej.



Rys. 1. Dendrogram hierarchicznej analizy skupień metodą Warda

Źródło: opracowanie własne na podstawie danych z BDL i GUS.

Do oszacowania odległości pomiędzy skupieniami wykorzystano metodę Warda. Metoda ta różni się od pozostałych, ponieważ wykorzystuje podejście analizy wariancji, dąży do minimalizacji sumy kwadratów odchyleń wewnątrz skupień. Metoda Warda uznawana jest za efektywną, chociaż jej stosowanie zmierza do tworzenia skupień o małej wielkości [Stanisz 2007]. Wynikiem badania jest dendrogram, czyli graficzna interpretacja uzyskanych wyników.

Istotnym elementem analizy skupień jest odcięcie dendrogramu, pozwalające określić liczbę skupień w analizowanym badaniu. W przeprowadzonej analizie ustalono wartość krytyczną na dwa sposoby. Pierwszy sposób wykorzystuje regułę R. Mojeny (1997) opisaną poniższym wzorem:

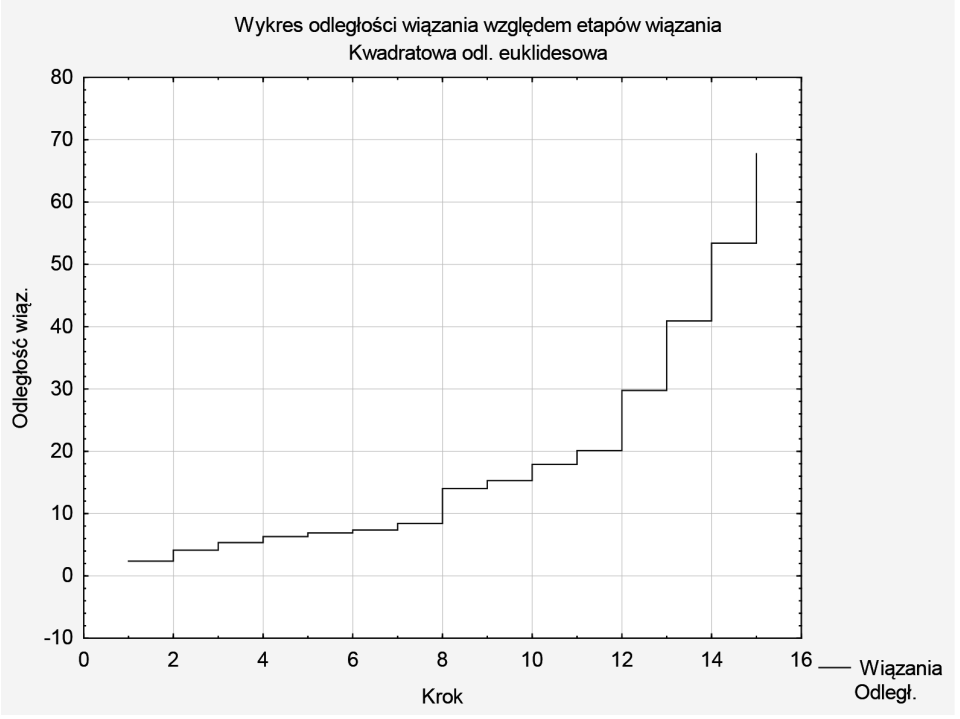
$$\hat{d}_{h+1} > \bar{d} + k\,S(d). \tag{3}$$

Uwzględniając $\bar{d}$ (średnią arytmetyczną długości wiązań) i $S(d)$ (odchylenie standardowe długości wiązań), R. Mojena proponuje przyjąć wartość parametru $k \in (2{,}75;\ 3{,}50)$. W praktyce najczęściej stosowaną wielkością parametru jest $k = 1{,}25$ określone na podstawie badań Milligan i Cooper [1985]. Alternatywnie zamiast k możemy przyjąć wartość z rozkładu t-Studenta, co oznacza dodatkowe założenie, że poziom połączenia klas na wykresie drzewa podlega rozkładowi normalnemu [Mikulec 2012]. Przyjmując $k = 1{,}25$, odcięcie następuje dla $\hat{d}_{h+1} > 44{,}5926$.

Korzystając z reguły R. Mojeny, otrzymamy trzy skupienia. Do pierwszego skupienia należą województwa: kujawsko-pomorskie, opolskie, warmińsko-mazurskie, zachodniopomorskie, lubuskie, świętokrzyskie i podkarpackie. Odrębne skupienie stanowi województwo mazowieckie. Jest to element odstający w badanej zbiorowości. Ostatnie, trzecie skupienie stanowią województwa: lubelskie, dolnośląskie, małopolskie, pomorskie, wielkopolskie, podlaskie, śląskie i łódzkie.

Drugim sposobem ustalenia wartości krytycznej jest analiza wykresu przebiegu aglomeracji, wykresu liniowego odległości wiązań względem kolejnych etapów procesu wiązania. Po zaobserwowaniu największego przyrostu, w którym tworzy się wiele skupień w przybliżeniu w takiej samej odległości wiązania, następuje odcięcie dzielące zbiór na klasy.

Na podstawie rysunku 2 można stwierdzić, że punkt odcięcia położony jest pomiędzy krokiem 8 a 9. Stąd możliwe jest wyodrębnienie siedmiu skupień. Pierwsze skupienie, to województwa kujawsko-pomorskie, opolskie, warmińsko-mazurskie i zachodniopomorskie. Skupienie drugie stanowi województwo lubuskie. W skupieniu trzecim znajdują się województwa świętokrzyskie i podkarpackie. Odrębnymi skupieniami (czwartym i piątym) są województwa mazowieckie i lubelskie. Do skupienia szóstego należy województwo: dolnośląskie i małopolskie. Ostatnie siódme skupienie, to województwa: pomorskie, wielkopolskie, podlaskie, śląskie i łódzkie.



Rys. 2. Wykres przebiegu aglomeracji

Źródło: opracowanie własne na podstawie danych BDL i GUS.

4. Porządkowanie liniowe

Zastosowanie taksonomicznego miernika Hellwiga [1968] pozwala ustalić klasyfikację obiektów (województw). Na podstawie standaryzowanych zmiennych wyznaczono wzorzec i antywzorzec, do których porównano badane województwa. Wzorzec i antywzorzec to abstrakcyjne obiekty o najlepszych i najgorszych wartościach cech. Wzorzec wyznaczamy poprzez obliczenie wartości maksymalnej dla stymulanty oraz wartości minimalnej dla destymulanty. Odwrotnie przyjmujemy funkcję maksimum i minimum dla antywzorca. Koncepcja ta rozwijana oraz stosowana była w wielu pracach [Pietrzak 2014]. Badane cechy, poza cechami X_7 (udział bezrobotnych zarejestrowanych z wyższym wykształceniem w liczbie ludności w wieku produkcyjnym), X_9 (liczba studentów przypadająca na jednego nauczyciela akademickiego) będącymi destymulantami, stanowią stymulanty. W kolejnym kroku obliczamy odległość każdego obiektu od wzorca, stosując metrykę euklidesową przedstawioną wzorem 4:

$$d_{i0} = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}. \quad (4)$$

Im bardziej obiekt (x_i) jest podobny do wzorca (y_i), tym wyższy jest poziom zjawiska złożonego dla tego obiektu.

Końcowym etapem badania jest wyznaczenie miary rozwoju przedstawionej wzorem

$$m_i = 1 - \frac{d_{i0}}{d_0}, \quad (5)$$

gdzie: d_0 – odległość pomiędzy wzorcem i antywzorcem, obliczona na podstawie metryki euklidesowej.

W analizowanym badaniu d_0 = 12,14. Im wyższe wartości m_i, tym wyższy poziom złożonego zjawiska. Miara rozwoju dla wzorca wynosi 1, zaś dla antywzorca 0, zatem $m_i \in [0, 1]$.

Tabela 2. Wyniki porządkowania liniowego

Pozycja w rankingu	Województwo	d_{i0}	m_i
1	mazowieckie	4,79	0,61
2	dolnośląskie	6,03	0,50
3	pomorskie	6,37	0,48
4	małopolskie	6,56	0,46
5	wielkopolskie	6,89	0,43
6	łódzkie	7,15	0,41
7	zachodniopomorskie	7,24	0,40
8	lubelskie	7,33	0,39
9	podlaskie	7,76	0,36
10	opolskie	7,86	0,35
11	śląskie	8,02	0,34
12	kujawsko-pomorskie	8,21	0,32
13	podkarpackie	8,77	0,28
14	warmińsko-mazurskie	8,79	0,27
15	lubuskie	8,99	0,26
16	świętokrzyskie	9,77	0,20

Źródło: opracowanie własne na podstawie danych BDL i GUS.

Analizując otrzymane wyniki, możemy podzielić obiekty na trzy klasy, w zależności od wielkości miary m_i. „Najlepsze" obiekty to województwa, dla których zachodzi relacja: $m_i \geq m_r$, klasa obiektów osiągających „najniższe" rezultaty to grupa województw spełniających relację $m_i \leq m_s$ [Paliszkiewicz 2010]. Przedział (m_s, m_r) przedstawiony wzorem 6 wynosi odpowiednio (0,276;0,483):

$$(m_s, m_r) = \left(\overline{m}_i - S(m_i), \overline{m}_i + S(m_i)\right), \tag{6}$$

gdzie: $\overline{m}_i$ – średnia arytmetyczna miary m_i, $S(m_i)$ – odchylenie standardowe miary m_i.

Uwzględniając powyższe warunki, możemy podzielić województwa na grupę województw, w których stan szkolnictwa wyższego jest „najlepszy": mazowieckie, dolnośląskie i pomorskie, grupę województw „przeciętnych": małopolskie, wielkopolskie, łódzkie, zachodniopomorskie, lubelskie, podlaskie, opolskie, śląskie, kujawsko-pomorskie, oraz grupę województw osiągających „najniższe rezultaty": podkarpackie, warmińsko-mazurskie, lubuskie i świętokrzyskie.

5. Zakończenie

Zaprezentowane metody – analiza skupień i porządkowania liniowego – w podobny sposób grupują województwa w odniesieniu do stanu szkolnictwa wyższego w Polsce. Warto zauważyć, że województwo mazowieckie, stanowiące odrębne skupienie w analizie skupień zajmuje pierwszą pozycję w rankingu utworzonym przy pomocy metod porządkowania liniowego. Analiza taksonomiczna potwierdziła zróżnicowanie potencjału edukacyjnego, wysoką pozycję województwa mazowieckiego oraz niską pozycję województw podkarpackiego, warmińsko-mazurskiego, lubuskiego i świętokrzyskiego. Podobne wyniki uzyskano w analizie szkolnictwa wyższego w latach 1999-2005 [Kwiatkowski, Roszkowska 2008].

Uwzględniając podział według reguły R. Mojeny, skupienie pierwsze składa się z województw „przeciętnych" i „najlepszych", zaś wszystkie województwa osiągające „najniższe" rezultaty (świętokrzyskie, lubuskie, podkarpackie, warmińsko-mazurskie) należą do wspólnej grupy – skupienia trzeciego.

Analizując podział według wykresu przebiegu aglomeracji, możemy spodziewać się innych pozycji w rankingu dla województw zachodniopomorskiego i śląskiego (pozycja w rankingu odpowiednio 7 i 11). Województwo zachodniopomorskie, zajmujące pozycję 7 w rankingu, w skupieniu otoczone jest województwami zajmującymi pozycje 10-14. Natomiast województwo śląskie (pozycja 11 w rankingu) należy do jednego skupienia m.in. z województwem pomorskim, które uzyskało 3 pozycję w rankingu.

Miara rozwoju dla województw z pozycją 1 i 2 różni się o ok. 0,1. Istotna różnica pojawia się również na pozycjach 12 i 13 (wynosi odpowiednio 0,32 i 0,28). Warto zauważyć, że „najlepsze" województwo (mazowieckie) uzyskało miarę rozwoju na

poziomie 0,61, zaś miara rozwoju dla wzorca jest równa 1, co może świadczyć, że nadal należy doskonalić jakość kształcenia na polskich uczelniach wyższych.

Literatura

Gatnar E., Walesiak M., 2004, *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej, Wrocław.

GUS, 2014, *Szkoły wyższe i ich finanse w roku 2014*, http://stat.gov.pl/obszary-tematyczne/edukacja/edukacja/szkoly-wyzsze-i-ich-finanse-w-2014-r-,2,11.html# (10.10.2016).

GUS, 2016, Bank Danych Lokalnych, https://bdl.stat.gov.pl/BDL/start (10.10.2016).

Hellwig Z., 1968, *Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom rozwoju i strukturę kwalifikowanych kadr*, Przegląd Statystyczny, nr 4, s. 307-327.

http://ncn.gov.pl/sites/default/files/pliki/NCN_statystyki_2014_pl.pdf.

Kwiatkowski E., Roszkowska S., 2008, *Rozwój i zróżnicowanie regionalne szkolnictwa wyższego w Polsce*, Gospodarka Narodowa, nr 4, s. 1-20.

Milligan G.W., Cooper M.C., 1985, *An examination of procedures for determining the number of clusters in a data set*, Psychometrika, vol. 50(2), s. 159-179.

Mikulec A., 2012, *Metody oceny wyniku grupowania w analizie skupień*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 242, Taksonomia 19, s. 460-468.

Paliszkiewicz J.O., 2010, *Wykorzystanie metody wzorca rozwoju do klasyfikacji przedsiębiorstw pod względem poziomu zarządzania wiedzą*, [w:] Knosala R. (red.), *Komputerowo zintegrowane zarządzanie*, t. 2, Oficyna Wydawnicza PTZP, Opole, s. 344-350.

Pietrzak M., 2014, *Taksonomiczny miernik rozwoju (TMR) z uwzględnieniem zależności przestrzennych*, Przegląd Statystyczny, nr 2, s. 181-201.

Stanisz A., 2007, *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. Tom 3. Analizy wielowymiarowe*, StatSoft, Kraków.

Strzała K., Przechlewski T., 1994, *Ekonometria inaczej*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.