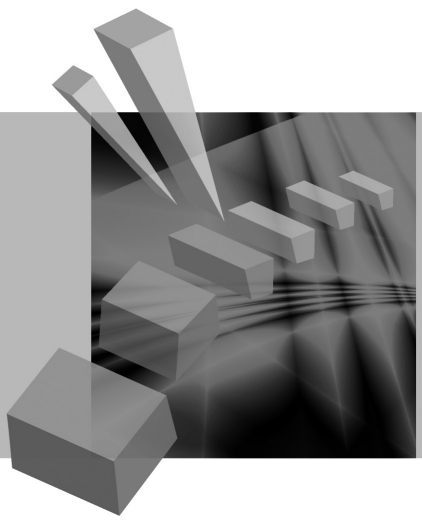
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana
Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Danuta Tarka
Politechnika Białostocka

WPŁYW METODY DOBORU CECH DIAGNOSTYCZNYCH NA WYNIKI KLASYFIKACJI OBIEKTÓW NA PRZYKŁADZIE DANYCH DOTYCZĄCYCH OCHRONY ŚRODOWISKA

Streszczenie: Celem pracy jest analiza wpływu sposobu doboru cech diagnostycznych na wyniki klasyfikacji obiektów w badaniach regionalnych na przykładzie danych o stanie środowiska w poszczególnych województwach. W badaniu użyto sześciu metod podstawowych w różnych wersjach (metody Hellwiga, medianowej metody Hellwiga, metody prostej grafowej, *k*-średnich, Warda, analizy czynnikowej i Bartosiewicz doboru reprezentantek). Dało to łącznie z wyjściowym $n = 18$ zbiorów cech, na których dokonano rangowania obiektów. Porównanie wyników rangowania z użyciem metod doboru i z rangowaniem opartym na zbiorze wyjściowym pokazało, że trzy z metod można uznać za substytucyjne wobec siebie i wobec rangowania bez doboru cech. Pozostałe wyniki rangowania były znacząco różne od rangowania bez eliminacji cech.

Słowa kluczowe: dobór cech diagnostycznych, klasyfikacja, porządkowanie liniowe.

1. Wstęp

W badaniach typu taksonomicznego wyróżnia się zwykle kilka etapów. Pierwszy z nich to określenie celu i zakresu badania i jest to etap specyficzny dla każdego badania. Następnym etapem jest dobór cech do badania. Nie jest to jednak problem zbyt często podejmowany w badaniach praktycznych i nadal nie wypracowano „(…) efektywnej procedury statystycznej wyboru cech" [Wysocki 2010, s. 45]. Dobór cech do badania dzieli się, zasadniczo, na dwa etapy: merytoryczny i formalny. W większości badań empirycznych autorzy podają zestaw dobranych zmiennych bez szerszej dyskusji sposobu doboru. Tymczasem już na etapie analizy merytorycznej, czyli określania, jakie własności powinny mieć cechy diagnostyczne, nie ma pełnej zgodności wśród autorów[1]. Etap doboru formalnego jest jeszcze słabiej reprezentowany w literaturze[2].

[1] Przegląd tych dyskusji autorka przedstawiła w pracy [Tarka 2010].

[2] W polskiej literaturze szerzej zajmuje się tym problemem Walesiak (np. [2005]), podobnie jednak jak inni autorzy, np. D. Steinly, M.J. Brusco [2008], analizuje problem dla zmiennych stochastycznych

Wielu autorów uważa, że wystarczy tylko uwzględnienie postulatu dyskryminacji cech, co sprowadzają do użycia współczynnika zmienności do eliminacji cech. Na ile jest to jednak wystarczające narzędzie, jest pytaniem do dyskusji. Jednym z argumentów za więcej niż tylko merytoryczną analizą cech, związaną z dziedziną i zakresem badania, jest postulat, by do porządkowania obiektów dobierać cechy zgodnie z ogólną zasadą: maksymalny zasób informacji przy minimalnej liczbie cech.

Pytanie podstawowe zadane w tej pracy brzmi: w jaki stopniu metody doboru cech wpływają na wyniki klasyfikacji obiektów. Celem precyzyjnego określenia kryterium porównania zawężono badanie do analizy wyników porządkowania liniowego opartego na cechach dobranych różnymi metodami, a co za tym idzie – do analizy *podobieństwa uporządkowania obiektów*, a nie podobieństwa uzyskanych zbiorów cech[3].

Jeżeli obszar badania jest wąski lub dostępna jest mała liczba cech, do kilkunastu, wówczas nie ma, w zasadzie, problemu z doborem cech do badania. Badacz uwzględnia, na ogół, wszystkie dostępne cechy. Dobór cech zaczyna być istotny, gdy np. cel badania jest określony szeroko (np. poziom rozwoju społecznego itp.), a podstawy teoretyczne zjawiska wymagają do poprawnej analizy dużego zbioru cech. Analiza merytoryczna, choć nadrzędna, może okazać się niewystarczająca.

W niniejszym badaniu zastosowano dwie metody do całego zbioru potencjalnych cech diagnostycznych (metodę Hellwiga i jej medianową odmianę), otrzymując zestaw cech reprezentujących badane zjawisko jako całość , analizę czynnikową oraz trzy metody reprezentujące podejście dualne[4].

Celem ujednolicenia sposobu uzyskania wyników tak, by tylko typ użytej metody doboru różnicował wynik uporządkowania obiektów, przyjęto następujące założenia:

1) podstawowym punktem odniesienia do porównań będzie wynik rankingu uzyskany na zbiorze cech potencjalnych (bez użycia jakiejkolwiek metody doboru, ale po eliminacji cech o współczynniku zmienności poniżej 10%),

2) przy doborze cech, tam, gdzie to było niezbędne, przyjęto jako progową wartość współczynnika korelacji liniowej Pearsona na poziomie $r^* = 0{,}7$,

3) wszystkie otrzymane zbiory cech diagnostycznych standaryzowano i ujednolicono zmienne do postaci stymulant poprzez odwrotność,

---

od strony teoretycznej, tzn. zakłada typ rozkładu i za pomocą symulacji analizuje wyniki klasyfikacji. Podstawowym jednak problemem w badaniach realnych procesów społeczno-gospodarczych jest to, iż nie znamy rzeczywistych rozkładów zmiennych, oraz to, że mając do czynienia z cechami empirycznymi opisującymi całą zbiorowość nie możemy przyjąć założenia o typie rozkładu zmiennych, zwłaszcza że najpopularniejsze założenie o normalności rozkładu jest w przypadku cech społeczno-gospodarczych mocno wątpliwe do przyjęcia.

[3] Problem od tej strony przedstawiano np. w pracach [Hadasik1993; Nowak 1981].

[4] Polega na klasyfikacji cech w grupy cech podobnych, a następnie na wybraniu reprezentantek grup.

4) porządkowanie liniowe przeprowadzono, konstruując miarę syntetyczną metodą Hellwiga [1968],

5) do porównania wyników rankingów uzyskanych na podstawie poszczególnych zbiorów cech użyto współczynnika korelacji rang Spearmana,

6) tam, gdzie to było niezbędne, użyto odległości euklidesowej, miejskiej i korelacyjnej w postaci $d = 1 - |r|$.

2. Materiał statystyczny i wstępna eliminacja cech

Do analizy jako przykładowy przyjęto zbiór danych dotyczących stanu i ochrony środowiska w układzie wojewódzkim dla roku 2005. Przyjętym kryterium uporządkowania jest ocena stanu i ochrony środowiska w ujęciu wojewódzkim. Jest to kryterium wystarczająco szerokie, by mieć do dyspozycji duży i dosyć spójny merytorycznie[5] materiał wyjściowy.

Ten wyjściowy zbiór cech przekształcono do postaci wskaźników, uzyskując ostatecznie $k = 80$ cech potencjalnych. Następnie przeanalizowano go z użyciem dwóch „konkurencyjnych"[6] względnych miar dyspersji[7]. W efekcie do dalszej analizy przyjęto dwa zbiory cech potencjalnych oznaczone jako: Zb1– powstały na podstawie zastosowania klasycznego współczynnika V_s zawierającego $n = 74$ cechy oraz Zb2 – powstały w wyniku użycia współczynnika V_{MOB} zawierającego $n = 71$ cech. Na tym etapie sprawdzano, jak bardzo użyte miary dyspersji cech zróżnicują zbiór wyjściowy[8].

3. Metody doboru cech

W pierwszym kroku zastosowano do zbiorów Zb1 i Zb2, niezależnie, te same procedury doboru cech, uporządkowano obiekty i porównano podobieństwo uzyskanych rankingów. W ten sposób sprawdzono także, na ile użycie dwóch różnych współczynników zmienności zróżnicowało wyniki rankingów. Do porównania, przede wszystkim skutków wstępnego zastosowania różnych współczynników zmienności,

---

[5] Rok 2005 jest ostatnim, w którym opublikowano szczegółowe dane dotyczące zanieczyszczeń powietrza w ujęciu wojewódzkim. Z ponad 20 cech reprezentujących zanieczyszczenia emitowane przez wszystkie zakłady publikuje się osiem rodzajów dotyczących tylko emisji zanieczyszczeń z zakładów szczególnie uciążliwych. Województwa uporządkowano według miary syntetycznej reprezentującej stopień zanieczyszczenia środowiska oraz działalność na rzecz jego ochrony.

[6] W literaturze często kwestionuje się zasadność użycia średniej arytmetycznej jako punktu odniesienia, rekomendując w zamian medianę jako mniej wrażliwą na nietypowe wartości, jest to jednak sprawa dyskusyjna.

[7] Klasyczny współczynnik zmienności (V_s) i $V_{MOB} = \dfrac{Me\left|x_{ij} - Me(x_j)\right|}{Me(x_j)} = \dfrac{MOB}{Me(x_j)}$.

[8] W tym przypadku zastosowanie współczynnika V_{MOB} wyeliminowało cztery *dodatkowe* cechy w stosunku do V_s.

użyto[9] parametrycznej metody Hellwiga (MH) oraz jej medianowej modyfikacji[10] (MHme). Obie metody zastosowano do każdego ze zbiorów Zb1 i Zb2 w całości oraz do doboru cech w poszczególnych działach[11]. W tabeli 1 przedstawiono współczynniki korelacji rang pomiędzy wynikami otrzymanymi z zastosowaniem wymienionych metod doboru cech do zbiorów wyjściowych Zb1 i Zb2 oraz z zastosowaniem tylko współczynników zmienności (Zb1.c i Zb2.c), czyli bez doboru. Jak widać z tab. 1, pomiędzy rankingami obiektów opartymi na cechach dobranych przy zastosowaniu klasycznego współczynnika zmienności (Zb1) i medianowego odchylenia względnego (Zb2) nie ma dużej różnicy. Zbieżność uporządkowań jest rzędu r_s = 98,5%. Jednak zastosowanie różnych metod doboru powoduje zwiększenie różnic pomiędzy wynikami. Najmniejsze różnice występują przy zastosowaniu oryginalnej metody Hellwiga do całości zbiorów Zb1.c i Zb2.c niezależnie, r_s = 90,59%. Analogiczne użycie metody Hellwiga z medianą (M.Hme.c) znacznie zmniejsza podobieństwo rankingów pomiędzy obydwoma zbiorami (r_s = 77,35%).

Tabela 1. Korelacja rang Spearmana wyników rankingów pomiędzy zbiorem Zb1 i Zb2

Metoda doboru	**Zb1.c**	M.H.c*	M.Hme.c	M.H działy	MHme. działy
Zb2.c	0,9853				
M.H.c		0,9059			
M.Hme.c			0,7735		
M.H. działy				0,5971	
MHme. działy					0,8676

* Litera **c** przy skrócie nazwy metody oznacza, że metodę stosowano do danego zbioru cech potencjalnych jako **c**ałości.

Źródło: obliczenia własne na podstawie danych GUS.

Bardzo dużą rozbieżność pomiędzy rankingami obserwujemy, gdy metoda Hellwiga (M.H. działy) jest stosowana do każdego działu w poszczególnych zbiorach jako całościach, odrębnie. Korelacja pomiędzy rankingami na poziomie r_s = 59,7% sygnalizuje dużą rozbieżność wyników. W przypadku użycia mediany do doboru

[9] W prezentowanej tu części badania nie uwzględniono metody macierzy odwrotnej, ponieważ odwrotne macierze współczynników korelacji dla Zb1 i Zb2 były źle uwarunkowane, dając na całej przekątnej liczby znacząco większe od 10.

[10] Zob. [Hellwig 1981]; przypomnijmy – modyfikacja polega na użyciu mediany zamiast średniej arytmetycznej, zob. np. [Młodak 2006].

[11] Cechy w obu zbiorach zostały przypisane do 7 poszczególnych działów merytorycznych (ziemia, woda itd.), tak ja są umieszczone w roczniku (a więc grupowanie cech było „eksperckie", z góry ustalone), a następnie do każdego działu zastosowano metodę doboru cech, uzyskując cechy reprezentujące poszczególne działy. Otrzymane podzbiory cech scalano jako jeden zbiór i utworzono na jego podstawie ranking województw, niezależnie dla Zb1 i Zb2.

cech w poszczególnych działach wyniki pomiędzy zbiorami są znacznie bardziej zbliżone. Można też zauważyć, że medianowa metoda Hellwiga zastosowana do działów (r_s = 86,76%) dała bardziej zbliżone wyniki na obu zbiorach niż zastosowana do zbiorów cech jako całości (r_s = 77,35%).

W następnym etapie zastosowano dodatkowe metody doboru cech już tylko na zbiorze Zb1, eliminując w ten sposób wpływ typu użytego uprzednio współczynnika zmienności. Były to: *analiza czynnikowa*[12]*, metoda grafowa, metoda Warda, metoda k-średnich*.

Schemat postępowania i liczbę pozostawionych przez daną metodę cech przedstawiono[13] na rys. 1.

Pozostałe, poza analizą czynnikową, metody użyte na Zb1 są *metodami dualnymi* bezpośrednio użytymi do pogrupowania obiektów. Do wyboru reprezentantki grup użyto jednej metody – Bartosiewicz[14] z modyfikacją Nowaka [1990, s. 57]. Jako reprezentantkę grupy wybierano cechę o najmniejszej średniej odległości korelacyjnej od pozostałych cech w grupie[15]. W przypadku metod aglomeracyjnych kluczowymi decyzjami w badaniu było określenie sposobu liczenia odległości pomiędzy cechami w przypadku metody Warda oraz sposobu określenia wstępnych centrów skupień w *k*-średnich[16]. Obu metod użyto raz do klasyfikacji cech w postaci wyjściowych wskaźników, tak jak to jest wspominane w literaturze[17], drugi raz, wprowadzając klasyfikację na podstawie odległości korelacyjnej[18]. Przyjęto zasadę, że podział powinien być zbliżony do liczby grup eksperckich (siedem) i otrzymano podział na $k = 9$, a dla odległości korelacyjnej $k = 8$ grup w obu metodach.

---

[12] Jest jedną z często rekomendowanych metod redukcji zbioru cech, a wartości pierwszej składowej uznaje się za miarę syntetyczną umożliwiającą uporządkowanie obiektów. Biorąc jednak pod uwagę wymagania stawiane cechom mającym służyć klasyfikacji i/lub porządkowaniu obiektów, można mieć duże zastrzeżenia co do zasadności użycia tej metody w najczęściej proponowany sposób. Procedura nie spełnia warunków reprezentatywności i co najwyżej słabego skorelowania użytych cech.

[13] Brakujące oznaczenia są objaśnione w przypisach 16, 17, 19.

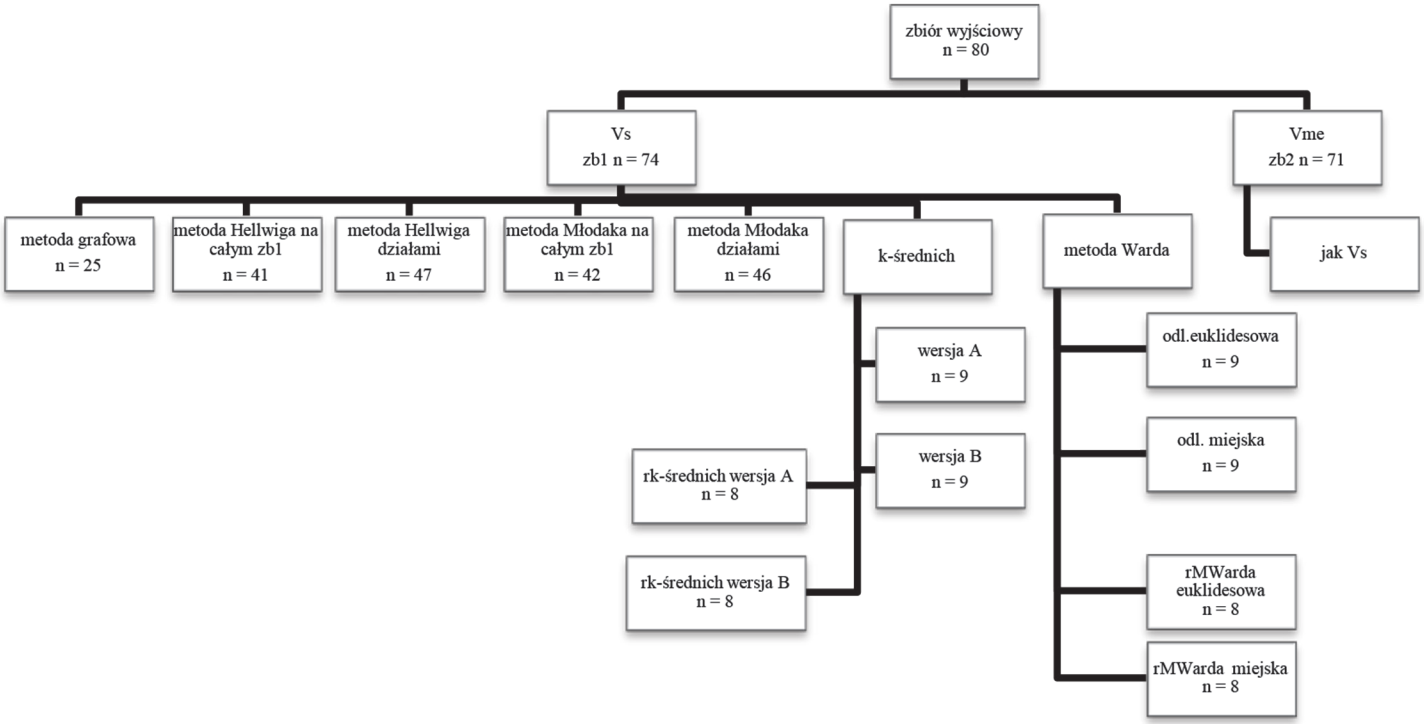
[14] Dokładniejszy przegląd metod grafowych można znaleźć np. w pracy [Grabiński i in. 1982, s. 138-145].

[15] W przypadku klasyfikacji metodą grafową nie udało się jednoznacznie rozstrzygnąć czy reprezentantką jednej z grup ma zostać cecha X19 czy X68, wobec czego przyjęto dwa odrębne zbiory cech diagnostycznych. W jednym pozostała pierwsza z tych cech (zbiór graf19), w drugim zaś druga cecha (zbiór graf68).

[16] Dla metody Warda przyjęto dwie metryki euklidesową oraz miejską. Podobnie przy *k*-średnich, w wersji a przyjęto maksymalizację odległości skupień, w wersji b zaś – sortowanie odległości i branie obserwacji przy stałym interwale. Nazwy metod w skrótach odpowiednio: Ward eukl, Ward miejska, k-śred **a**, k-śred **b**.

[17] Użyto tego słowa celowo, na ogół bowiem autorzy przy omawianiu metod klasyfikacji obiektów wspominają, że można ich użyć do klasyfikacji cech bez wgłębiania się w szczegóły ich stosowania, zob. np. [Gan, Ma, Wu 2007; Grabiński i in. 1982].

[18] W tabeli będą one oznaczone jako **r**Ward eukl, **r**Ward miejska, **r**k-śred **a**, **r**k-śred **b.**



Rys. 1. Użyte metody doboru cech diagnostycznych i liczba uzyskanych cech (*n*)

Źródło: opracowanie własne.

4. Wyniki badania

W wyniku użycia wymienionych metod doboru wraz z ich odmianami otrzymano **m = 14** zbiorów cech diagnostycznych wyodrębnionych już tylko na podstawie zbioru Zb1, a następnie dokonano na tych zbiorach uporządkowania obiektów. Do porównania stopnia podobieństwa rankingów między nimi użyto współczynnika korelacji rang Spearmana. Wartości współczynników korelacji rang Spearmana dla wszystkich rankingów przedstawiono w tab. 2. Na jej podstawie można przeanalizować stopień podobieństwa rankingów między poszczególnymi metodami. Zaczynając analizę od spojrzenia na podobieństwo wyników rangowania tymi samymi metodami różniącymi się tylko: sposobem stosowania[19], określania podobieństwa[20] punktem odniesienia[21], możemy stwierdzić, że już przy tym najprostszym różnicowaniu metod wyniki rankingów odbiegają od siebie znacząco. Zbliżone uporządkowania otrzymano dla klasycznej i medianowej metody Hellwiga zastosowanych do działów merytorycznych ($r_s = 0{,}903$) oraz do całości zbioru ($r_s = 0{,}871$). Jeszcze tylko metoda *k*-średnich w wersji a i b[22] daje podobieństwo wyników na poziomie $r_s > 0{,}7$. Pozostałe metody dają bardzo różniące się uporządkowania obiektów. Przechodząc do porównania współczynników korelacji rang różnych metod parami, możemy stwierdzić, że największe podobieństwo między uporządkowaniami wykazują wyniki uzyskane metodą grafową i klasyczną metodą Hellwiga dla całości zbioru ($r_s = 0{,}821$) oraz między grafową i medianową metodą Hellwiga na całości zbioru. Jeśli badacz założy, że osiemdziesięcioprocentowe podobieństwo jest wystarczające, wówczas możemy uznać te trzy metody za substytucyjne pomiędzy sobą. Pozostałe metody wykazały podobieństwa rankingów pomiędzy sobą na poziomie poniżej $r_s < 0{,}7$, poza metodami *rk*-średnich i *r*Ward miejską ($r_s = 0{,}747$). Obie te metody mają jednak niskie współczynniki korelacji rang z uporządkowaniami z pozostałymi metodami. Najmniej podobne wyniki otrzymano pomiędzy rankingami metodą Warda i *k*-średnich we wszystkich ich odmianach a pozostałymi metodami.

Ostatnim krokiem analizy jest prześledzenie, jak bardzo zbliżone do siebie są rankingi z wykorzystaniem poszczególnych metod doboru cech i rankingu bez doboru. W tym celu przeanalizujemy pierwszą kolumnę tab. 2.

Punktem centralnym porównania jest ranking oparty na wyjściowym zbiorze Zb1, a więc bez doboru cech. *Pytanie badawcze jest następujące: na ile dobór cech (już choćby poprzez redukcję ich liczby) zmienia wyniki uporządkowania obiektów i na ile prezentowane metody są substytucyjne w stosunku do uporządkowania na*

[19] Do całości zbioru cech potencjalnych lub do jego działów merytorycznych odrębnie.
[20] Odległość miejska, euklidesowa, korelacyjna (metoda Warda, *k*-średnich).
[21] Średnia, mediana.
[22] Różnice zob. przypis 13.

Tabela 2. Korelacja rang Spearmana pomiędzy wynikami rankingów przy różnych metodach doboru cech dla zbioru Zb1

Metoda	**zb1 całość**	M.H.c	M.Hme.c	M.H. działy	M.Hme. działy	graf19	analiza czynni-kowa	k-średnich a	k-średnich b	Ward eukl	Ward miejska	rk-śred a	rk-śred b	rWard eukl	rWard miejska
zb1 całość	1														
M.H.c	**0,821**	1													
M.Hme.c	**0,832**	*0,871*	1												
M.H. działy	0,485	0,626	*0,779*	1											
M.Hme. działy	0,418	0,541	0,647	*0,903*	1										
graf19	**0,918**	*0,821*	*0,803*	0,444	0,294	1									
analiza czynnikowa	0,662	0,418	0,353	–0,135	–0,209	0,638	1								
k-średnich a	0,541	0,147	0,324	0,115	0,085	0,359	0,347	1							
k-średnich b	**0,803**	0,524	0,609	0,262	0,203	*0,703*	0,444	*0,706*	1						
Ward eukl	0,176	0,206	0,097	0,024	0,262	–0,029	0,026	0,185	0,147	1					
Ward miejska	0,176	0,124	0,344	0,485	0,671	–0,059	–0,197	0,338	0,276	0,471	1				
rk-śred a	0,774	0,521	0,612	0,300	0,291	0,674	0,456	0,588	*0,782*	0,132	0,194	1			
rk-śred b	0,485	0,432	0,524	0,235	0,071	0,515	0,129	0,129	0,632	–0,018	–0,059	0,644	1		
rWard eukl	0,397	0,453	0,515	0,518	0,574	0,244	–0,179	0,121	0,482	0,474	0,426	0,438	0,603	1	
rWard miejska	0,412	0,176	0,238	0,038	0,076	0,400	0,171	0,329	0,438	0,176	–0,109	*0,747*	0,559	0,350	1

Źródło: obliczenia własne na podstawie danych GUS.

zbiorze bez doboru cech[23]*?* Nie określimy, który ranking jest lepszy, bowiem nie mamy kryterium "lepszości[24]".

Pewnym zaskoczeniem okazało się, iż największe podobieństwo w porządkowaniu obiektów w stosunku do wyników zbioru wyjściowego Zb1 dała prosta metoda grafowa[25]. Podobieństwo wyników jest bardzo duże, zważywszy na to, iż zredukowała ona zbiór z $n = 74$ do $n = 25$ cech. Następny najbliższy wynik dała medianowa wersja metody Hellwiga, a za nią uplasowała się oryginalna metoda Hellwiga. Zakładając, że np. $r_s = 0{,}8$ jest progowym poziomem podobieństwa, w granicy tej mieści się jeszcze metoda *k*-średnich dla wskaźników w poziomach w wersji **b**[26]. Sporym zaskoczeniem okazało się małe podobieństwo wyników metody Hellwiga i medianowej w ujęciu działowym, co wymaga jednak dodatkowej analizy merytorycznej. Relatywnie dobrze też, biorąc pod uwagę przedstawione wcześniej zastrzeżenia, wypadła analiza czynnikowa, dając bardziej podobny wynik rankingu do rankingu na całości zbioru ($r_s = 0{,}662$) niż dwie wersje *k*-średnich i metoda Warda we wszystkich odmianach. Metoda Warda we wszystkich wersjach dała najmniej podobne wyniki i nie można tego wyjaśnić dużą redukcją liczby cech, bowiem identyczną liczbę cech miała *k*-średnich.

5. Wnioski

Biorąc pod uwagę tylko współczynniki zmienności jako sposób eliminacji cech[27], otrzymano małą różnicę ($r_s = 0{,}9853$) pomiędzy wynikami rankingów na alternatywnych zbiorach Zb1 i Zb2, co oznacza, że dodatkowa eliminacja cech znacząco zwiększyła zróżnicowanie wyników rankingów. Przy czym dobór cech dla zbioru jako całości dał wyższe współczynniki korelacji rang pomiędzy zbiorami Zb1 i Zb2 traktowanymi jako całości niż dobór cech dla każdego z działów odrębnie.

Podsumowując zaś wyniki uzyskane przy różnych metodach doboru na tym samym zbiorze wyjściowym (Zb1), można stwierdzić, że przyjmując jako progowe podobieństwo np. $r_s = 0{,}8$, możemy uznać metodę grafową, obie wersje metody

---

[23] W rozumieniu odpowiedzi na pytanie czy, w granicach zdefiniowanego przez badacza poziomu podobieństwa dają takie same uporządkowania.

[24] Określenie tego jest w badaniach społeczno-gospodarczych w dużej mierze elementem wiedzy merytorycznej, a nie formalno-statystycznej.

[25] Przy okazji można stwierdzić też, że cecha X19 lepiej reprezentuje swoją grupę niż X68, rozumiejąc przez to reprezentatywność informacyjną (nastąpiła mniejsza utrata informacji w stosunku do zbioru wyjściowego). Współczynnik korelacji rang pomiędzy wynikami z użyciem zbioru graf19 a zbiorem Zb1 bez doboru był wyższy niż zbioru graf68. Ten ostatni zbiór wyeliminowano więc z dalszego opisu.

[26] Przyjmowanej standardowo w obliczeniach.

[27] W tym przypadku użycie mediany nie zmieniło istotnie rangowania obiektów, choć cechy były mocno zróżnicowane pod względem siły i kierunku asymetrii.

Hellwiga oraz jedną z wersji *k*-średnich[28] za substytucyjne w stosunku do braku doboru cech. Wobec tego, w zależności od potrzeb, jeśli chcemy zredukować liczbę cech, zachowując maksymalną informację ze zbioru pierwotnego, wówczas można użyć którejś z powyższych metod[29].

Pozostałe metody i ich sposoby użycia dają na tyle różniące się wyniki, iż można uznać je za ,,konkurencyjne" w stosunku do wymienionej powyżej grupy i braku doboru cech. Powyższe wnioski obciążone są dużą niepewnością co do ich ogólności ze względu na to, że oparte są na jednym empirycznym przykładzie. Dalsze badania empiryczne oparte na danych dotyczących innych obszarów społeczno-gospodarczych pozwolą, być może, na większe uogólnienia. Jest to problematyka, która wymaga dalszych analiz.

Literatura

Gan G., Ma Ch., Wu J. (2007), *Data Clustering Theory, Algorithms, And Applications*, ASA-SIAM Series on Statistics and Applied Probability.

Grabiński T., Wydymus T., Zeliaś A. (1982), *Metody doboru zmiennych w modelach ekonometrycznych,* PWN, Warszawa.

Hadasik D. (1993), *Kilka uwag na temat porównywalności wyników różnych badań taksonomicznych*, „Przegląd Statystyczny", nr 2, s. 233-236.

Hellwig Z. (1968), *Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr*, „Przegląd Statystyczny", nr 4.

Hellwig Z. (1981), *Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych,* [w:] *Metody i modele matematyczno-ekonomiczne w doskonaleniu zarządzania gospodarką socjalistyczną*, red. W. Welfe, PWE, Warszawa.

Nowak E. (1981), *Badanie zgodności metod wyboru cech diagnostycznych*, „Przegląd Statystyczny", nr 3-4, s. 301-309.

Nowak E. (1990), *Problem informacji w modelowaniu ekonometrycznym*, PWN, Warszawa.

Ochrona środowiska 2006, (2006)*,* GUS, Warszawa, Informacje i Opracowania Statystyczne

Steinly D., Brusco M.J. (2008), *Selection of variables in cluster analysis: an empirical comparison of eight procedures*, „Psychometrika" vol. 73, no. 1, s. 125-144.

Tarka D. (2010), *Własności cech diagnostycznych w badaniach typu taksonomicznego*, Ekonomia i Zarządzanie, Politechnika Białostocka, Białystok, t. 2, nr 4, s. 194-205.

Walesiak M. (2005), *Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji,* [w:] K. Jajuga, M. Walesiak, *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 12, Prace Naukowe AE we Wrocławiu nr 1076, s. 106-118.

Wysocki F. (2010), *Metody taksonomiczne w rozpoznawaniu typów ekonomicznych rolnictwa i obszarów wiejskich*, Wyd. Uniwersytetu Przyrodniczego w Poznaniu, Poznań.

[28] Fakt, że jest to tylko jedna z wersji metody *k*-średnich, jest nieco kłopotliwy do wyjaśnienia i nie ma pewności, czy nie jest to wynik specyficzny związany z obszarem, dla którego przeprowadzano analizę.

[29] Bardziej precyzyjne wnioski można by wyciągnąć, powracając do analizy merytorycznej wyników, ale nie pozwala na to ograniczony zakres niniejszej pracy, analiza merytoryczne będzie więc przeprowadzona odrębnie.

INFLUENCE OF THE FEATURES SELECTION METHOD ON THE RESULTS OF OBJECTS CLASSIFICATION USING ENVIRONMENTAL DATA

Summary: The main aim of the paper is a comparison of the diagnostics features selection methods influence on regional objects linear and nonlinear classification using as an example official environmental data on Polish voivodeships. Six methods and their variations were used. Those used methods were: Hellwig method, median Hellwig method, simple graph method, k-means method, Ward method, factor analysis and Bartosiewicz method of finding group representative. Eighteen sets of diagnostic features were obtained which were used to rank objects. Spearman's coefficient of correlation was used to compare results of rankings. Three of the methods gave comparable results to original data ranking, the rest of them gave very different results.

Keywords: features selection methods, objects classification and ranking.