

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 327

Taksonomia 22

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowci

Krzysztof Jajuga, Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,

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

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2014

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

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Eugeniusz Gatnar , Balance of payments statistics and external competitiveness of Poland.....	15
Andrzej Sokolowski, Magdalena Czaja , Efektywność metody k -średnich w zależności od separowalności grup.....	23
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw	30
Elżbieta Gołata , Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011.....	39
Aleksandra Łuczak, Feliks Wysocki , Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów	49
Marek Walesiak , Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej	60
Paweł Lula , Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu <i>RAKE</i>	69
Mariusz Kubus , Propozycja modyfikacji metody złagodzonego LASSO.....	77
Andrzej Bąk, Tomasz Bartłomowicz , Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie <i>DiscreteChoice</i> programu R.....	85
Justyna Brzezińska , Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012.....	95
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Zastosowanie pakietu <i>dcMNM</i> programu R w badaniach preferencji konsumentów wódki	104
Barbara Batóg, Jacek Batóg , Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010	113
Małgorzata Markowska, Danuta Strahl , Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego.....	121
Kamila Migdał-Najman, Krzysztof Najman , Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	131
Kamila Migdał-Najman, Krzysztof Najman , Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena	139
Beata Basiura, Anna Czapkiewicz , Badanie jakości klasyfikacji szeregów czasowych	148
Michał Trzęsiok , Wybrane metody identyfikacji obserwacji oddalonych.....	157

Grażyna Dehnel, Tomasz Klimanek , Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej.....	167
Michał Bernard Pietrzak, Justyna Wilk , Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji.....	177
Maciej Beręsewicz , Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena	186
Marcin Szymkowiak, Tomasz Józefowski , Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji	195
Marcin Pelka , Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym	202
Małgorzata Machowska-Szewczyk , Ocena klas w rozmytej klasyfikacji obiektów symbolicznych.....	210
Justyna Wilk , Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych.....	220
Andrzej Dudek , Metody analizy skupień w klasyfikacji markerów map Google	229
Ewa Roszkowska , Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW	237
Marcin Szymkowiak, Marek Witkowski , Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych.....	248
Bartłomiej Jefmański , Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R.....	257
Karolina Bartos , Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych	266
Joanna Trzęsiok , Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych	275
Beata Bal-Domańska , Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja.....	285
Beata Bieszk-Stolorz, Iwona Markowicz , Wpływ zasiłku na proces poszukiwania pracy	294
Marta Dziechciarz-Duda, Klaudia Przybysz , Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni.....	303
Tomasz Klimanek , Problem pomiaru procesu dezagrarnizacji wsi polskiej w świetle wielowymiarowych metod statystycznych.....	313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Wybrane metody analizy danych wzdluznych.....	321
Artur Zaborski , Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych	330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy.....	338

Katarzyna Wawrzyniak , Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego	346
---	-----

Summaries

Eugeniusz Gatnar , Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski	22
Andrzej Sokółowski, Magdalena Czaja , Cluster separability and the effectiveness of k -means method	29
Barbara Pawelek, Józef Pocięcha, Adam Sagan , Multisectoral analysis of latent transitions in bankruptcy prediction models.....	38
Elżbieta Golata , Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011	48
Aleksandra Łuczak, Feliks Wysocki , Determination of weights for features in problems of linear ordering of objects	59
Marek Walesiak , Reinforcing measurement scale for ordinal data in multivariate statistical analysis	68
Paweł Lula , Automatic identification of keywords and keyphrases in documents written in Polish.....	76
Mariusz Kubus , The proposition of modification of the relaxed LASSO method.....	84
Andrzej Bąk, Tomasz Bartłomowicz , Microeconomic multinomial logit models and their implementation in the <code>DiscreteChoice</code> R package .	94
Justyna Brzezińska , The analysis of unemployment data in Poland in 2004-2012 with application of log-linear models	103
Andrzej Bąk, Marcin Pelka, Aneta Rybicka , Application of the MMLM package of R software for vodka consumers preference analysis.....	112
Barbara Batóg, Jacek Batóg , Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity	120
Małgorzata Markowska, Danuta Strahl , Classification of the European regional space in terms of smart growth pillars using the reference limit system.....	130
Kamila Migdał Najman, Krzysztof Najman , Formal quality assessment of group structure mapping on the Kohonen's map	138
Kamila Migdał Najman, Krzysztof Najman , Graphical quality assessment of group structure mapping on the Kohonen's map	147
Beata Basiura, Anna Czapkiewicz , Validation of time series clustering	156
Michał Trzęsiok , Selected methods for outlier detection.....	166

Grażyna Dehnel, Tomasz Klimanek , Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics	176
Michał Bernard Pietrzak, Justyna Wilk , Economic distance in modeling spatial phenomena with the application of gravity model.....	185
Maciej Beręsewicz , An attempt to use different distance measures in the Generalized Petersen estimator	194
Marcin Szymkowiak, Tomasz Józefowski , Construction and practical using of SPREE estimators for two-dimensional contingency tables.....	201
Marcin Pelka , The ensemble conceptual clustering for symbolic data.....	209
Małgorzata Machowska-Szewczyk , Evaluation of clusters obtained by fuzzy classification methods for symbolic objects.....	219
Justyna Wilk , Problem of determining the number of clusters in taxonomic analysis of symbolic data	228
Andrzej Dudek , Clustering techniques for Google maps markers.....	236
Ewa Roszkowska , The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure	247
Marcin Szymkowiak, Marek Witkowski , The use of correspondence analysis in analysing the financial situation of cooperative banks.....	256
Bartłomiej Jefmański , The construction of fuzzy customer satisfaction indexes using R program.....	265
Karolina Bartos , Discovering patterns of consumer behaviour by market basket analysis of the transactional data.....	274
Joanna Trzęsiok , Cluster analysis of countries with respect to fertility rate and other demographic factors	284
Beata Bal-Domańska , An attempt to identify major regional clusters and their convergence	293
Beata Bieszk-Stolorz, Iwona Markowicz , The influence of benefit on the job finding process	302
Marta Dziechciarz-Duda, Klaudia Przybysz , Education and labor market needs. Classification of university graduates	312
Tomasz Klimanek , The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods.....	320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska , Selected methods for an analysis of longitudinal data.....	329
Artur Zaborski , The application of distance measures for ordinal data for aggregation individual preferences	337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek , Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market	345
Katarzyna Wawrzyniak , Classification of voivodeships according to the level of the realization of priorities of <i>the National Development Strategy 2007-2015</i> with using the values of centroid of the rows	355

Justyna Wilk

Uniwersytet Ekonomiczny we Wrocławiu

PROBLEM WYBORU LICZBY KLAS W TAKSONOMICZNEJ ANALIZIE DANYCH SYMBOLICZNYCH

Streszczenie: W artykule rozważono problem wyboru liczby klas w analizie skupień obiektów symbolicznych. Zaproponowano procedurę, która ułatwi określenie struktury zbioru obiektów. Obejmuje ona ustalenie przedziału liczby klas, selekcję zmiennych symbolicznych z wykorzystaniem procedur formalnych, zastosowanie hierarchicznych i optymalizacyjnych metod klasyfikacji oraz ocenę wskazań indeksów wyboru liczby klas. Najlepsze wyniki uzyskano po wyeliminowaniu zmiennych zakłócających metodą *HINoV* oraz zastosowaniu metod Warda i DCLUST. Wskazania indeksów były jednak zróżnicowane w zależności od zastosowanej metody klasyfikacji i jej własności. Zatem ostateczna decyzja zależy od przyjętego celu badania.

Słowa kluczowe: dane symboliczne, analiza skupień, liczba klas, taksonomia numeryczna, taksonomia symboliczna.

1. Wstęp

Celem analizy skupień jest określenie struktury zbioru obiektów poprzez zgrupowanie w klasy obserwacji najbardziej do siebie podobnych. Nieodłącznym elementem analizy skupień, determinującym użyteczność otrzymanych wyników, jest wybór liczby klas. Metody taksonomiczne nie mają jednak wbudowanych procedur ustalania optymalnej liczby klas.

Problem wyboru liczby klas jest utrudniony w przypadku, gdy klasyfikacji podlegają obiekty opisane za pomocą danych symbolicznych (obiekty symboliczne). Dane symboliczne stanowią bardziej złożoną formę reprezentacji zjawisk niż dane w ujęciu klasycznym, wyrażane w postaci pojedynczej kategorii lub wartości liczbowej. Realizacjami zmiennych symbolicznych są przedziały wartości, zbiory kategorii, zbiory kategorii z wagami, prawdopodobieństwami i częstościami oraz struktury drzewiaste [zob. Bock, Diday (red.) 2000; Diday, Noirhomme-Fraiture (red.) 2008].

Mimo coraz większej złożoności problemów badawczych rozwiązywanych z wykorzystaniem metod taksonomicznych oraz rosnącego zainteresowania analizą

danych symbolicznych, problem wyboru liczby klas nie doczekał się wielu opracowań w literaturze przedmiotu. W artykule podjęto próbę sformułowania podejścia, które ułatwi wybór liczby klas w zbiorze obiektów symbolicznych.

2. Klasyfikacja obiektów symbolicznych

Dane symboliczne mogą wynikać ze specyfiki zjawiska bądź np. konstrukcji kwestionariusza ankiety. Jednostki badania podlegające klasyfikacji określane są wtedy jako obiekty symboliczne I rzędu. Dane symboliczne mogą być również rezultatem agregacji danych w ujęciu klasycznym. Przeprowadza się ją, gdy występuje konieczność redukcji opisu (w przypadku dużych zbiorów danych) bądź potrzeba uszczegółowienia opisu jednostek nadrzędnych. Takie jednostki badania określane są jako obiekty symboliczne II rzędu [zob. Bock, Diday (red.) 2000].

Klasyfikacja obiektów symbolicznych nie różni się od typowej procedury stosowanej dla danych w ujęciu klasycznym [zob. Punj, Stewart 1983, s. 144; Milligan 1996, s. 342-343; Walesiak 2004]. Jednak złożoność danych symbolicznych implikuje stosowanie specyficznych rozwiązań na każdym jej etapie, począwszy od selekcji obiektów i zmiennych, a skończywszy na opisie i profilowaniu klas [zob. Wilk 2010].

W klasyfikacji obiektów symbolicznych zastosowanie znajdują w szczególności metody taksonomii symbolicznej, np. metody SCLUST i DCLUST [zob. Bock, Diday (red.) 2000; Diday, Noirhomme-Fraiture (red.) 2008]. Stosowane są także metody taksonomii numerycznej, o ile bazują na macierzy odległości, np. Warda i k -medoidów [zob. Anderberg 1973; Everitt i in. 2001]. Przegląd zastosowań metod taksonomicznych w analizie danych symbolicznych przedstawiono w pracy [Wilk 2010].

3. Podejścia w wyborze liczby klas obiektów symbolicznych

Wybór liczby klas jest jednym z najtrudniejszych etapów analizy skupień i stanowi przedmiot rozważań wielu badaczy [zob. np. Milligan, Cooper 1985; Jain, Dubes 1998; Grabiński 1992, s. 101-102; Walesiak 2004, s. 338-341]. Na przykład w segmentacji rynku, z punktu widzenia celów przedsiębiorstwa, dąży się do podziału konsumentów na jak najbardziej homogeniczne grupy, aby lepiej dopasować do nich instrumenty marketingu mix. Implikuje to wybór większej liczby segmentów. Jednak należy mieć na uwadze możliwości finansowe i organizacyjne firmy z uwagi na tzw. korzyści skali. Segmenty powinny być na tyle duże, aby były atrakcyjne w perspektywie długookresowej. Z tego punktu widzenia dąży się do ograniczenia liczby segmentów [por. np. Wedel, Kamakura 1998, s. 60].

Wybór liczby klas dokonywany jest w oparciu o przesłanki merytoryczne (podstawy teoretyczne, wiedzę badacza, opinię ekspertów, wyniki wcześniejszych badań itd.). Jednak w sytuacji, gdy brakuje wystarczającej wiedzy, aby precyzyjnie

określić liczbę klas, stosuje się podejście merytorycznoformalne (z wykorzystaniem narzędzi statystycznych).

Zazwyczaj przyjmuje się następującą procedurę:

a) określa się przedział liczby klas *a priori* bądź z wykorzystaniem metod hierarchicznych i wskazań współczynnika aglomeracji, który informuje, o ile zmienił się poziom separowalności klas [zob. np. Hair i in. 2006],

b) stosuje się indeksy wyboru liczby klas [zob. np. Milligan, Cooper 1985] i poszukuje się ich wartości optymalnych oraz porównuje zgodność ich wskazań,

c) wybiera się podział, w którym klasy są unikatowe i znacznie się różnią, mają wysoką wartość poznawczą i użyteczność ze względu na cel badania.

Problem wyboru liczby klas obiektów symbolicznych był podejmowany m.in. w pracach [Gowda, Diday 1994; Lechevallier (red.) 2001; Hardy, Lallemand 2002; Verde i in. 2003; Hardy 2005; Diday, Noirhomme-Fraiture (red.) 2008]. W wyborze liczby klas obiektów symbolicznych zastosowanie mają indeksy:

- bazujące na tablicy danych symbolicznych, np. test *hypervolumes*, test *gap* Rassona i Kubushishi, indeks $Q(P)$ Verde, Lechevalliera i Chavent;
- bazujące na macierzy odległości, np. indeksy Bakera i Huberta, Huberta i Levine, *silhouette* Rousseeuwa, statystyka Huberta, *CI* Gowdy i Didaya. Ich zastosowanie wymaga pomiaru odległości za pomocą miar opracowanych w ramach analizy danych symbolicznych [zob. Bock, Diday (red.) 2000, s. 153-185; Malerba i in. 2001; Malerba i in. 2002, s. 33-35; Wilk 2005, 2006];
- bazujące na macierzy danych, np. indeks Calińskiego i Harabasza, Krzanowskiego i Lai, Hartigana, Daviesa-Bouldina, indeks *gap* Tibshirani, Walthera i Hastie, przy czym zamiast centroidów należy wyznaczyć medoidy (obiekty reprezentujące klasy, dla których suma odległości od pozostałych obiektów z klasy jest najmniejsza).

Analizy porównawcze indeksów wyboru liczby klas obiektów symbolicznych zawierają prace [Hardy, Lallemand 2002; Mali, Mitra 2003; Dudek 2007; Hardy 2005]. Przeprowadzone badania różnią się rodzajem danych (generowane, rzeczywiste) i wielkością zbiorów obiektów, rodzajem i liczbą zmiennych symbolicznych, sposobem postępowania z danymi symbolicznymi (dane pierwotne, dane przekształcone), rozważanymi indeksami wyboru liczby klas, zastosowanymi metodami klasyfikacji i miarami odległości oraz podejściem w interpretacji wyników (poziom użyteczności klas, kryterium separowalności i spójności, ocena zgodności wskazań indeksów, ocena zgodności z naturalnym podziałem).

W przypadku danych generowanych indeksy dawały zazwyczaj jednoznaczne wskazania (bez względu na metodę klasyfikacji) i podział zgodny z zadaną strukturą klas [zob. Dudek 2007; Hardy 2005]. Natomiast dla danych rzeczywistych wskazania indeksów były zróżnicowane, a wybór liczby klas znacznie trudniejszy. Powodem tego mógł być brak przeprowadzenia selekcji zmiennych w celu wyeliminowania zmiennych zakłócających strukturę klas. Stanowiło to motywację do podjęcia badań w tym zakresie.

4. Określenie liczby klas obiektów opisanych zmiennymi symbolicznymi

Na podstawie informacji ze stron internetowych autoryzowanych dealerów zgromadzono dane charakteryzujące wybrane modele samochodów osobowych, dotyczące gabarytów, osiągow, parametrów technicznych i cen. Uwzględniono samochody kwalifikowane do czterech autosegmentów, oznaczonych A, B, C i D (tab. 1). Przy-

Tabela 1. Zbiór obiektów symbolicznych

Lp.	Marka	Model	Auto-segment	Lp.	Marka	Model	Auto-segment
1	Skoda	Nowa Fabia	B	16	Opel	Astra	C
2	Skoda	Nowa Octavia	C	17	Volkswagen	Nowe Polo	B
3	Fiat	Panda	A	18	Volkswagen	Golf	C
4	Fiat	Grande Punto	B	19	Volkswagen	Passat Limousine	D
5	Fiat	Bravo	C	20	Chevrolet	Nowy Spark	A
6	Peugeot	308	C	21	Chevrolet	Aveo	B
7	Peugeot	407	D	22	Chevrolet	Lacetti	C
8	Citroen	C1	A	23	Seat	Ibiza	B
9	Citroen	Nowy C3	B	24	Seat	Leon	C
10	Citroen	C4	C	25	Seat	Exeo	D
11	Toyota	Aygo	A	26	Honda	Jazz	B
12	Toyota	Yaris	B	27	Honda	Civic 5D	C
13	Toyota	Corolla	C	28	Honda	Accord Sedan	D
14	Toyota	Avensis	D	29	Nissan	Micra	A
15	Opel	Corsa	B	30	Nissan	Tiida	B

Źródło: opracowanie własne.

Tabela 2. Zbiór zmiennych symbolicznych

Lp.	Nazwa zmiennej symbolicznej	Rodzaj zmiennej symbolicznej	Zbiór realizacji zmiennej symbolicznej	Jednostka miary
1	Cena katalogowa	przedział liczbowy	[27 990; 144 500]	zł
2	Typ nadwozia	lista kategorii	{hatchback; sedan; combi}	–
3	Rozstaw osi	przedział liczbowy	[2299; 2725]	mm
4	Długość nadwozia	przedział liczbowy	[3415; 4765]	mm
5	Szerokość nadwozia	przedział liczbowy	[1465; 2033]	mm
6	Wysokość nadwozia	przedział liczbowy	[1430; 1760]	mm
7	Pojemność skokowa	lista kategorii	{1,0; 1,1; 1,2; 1,3; 1,4; 1,6; 1,7; 1,8; 1,9; 2,0; 2,2; 2,4}	–
8	Moc silnika	przedział liczbowy	[54; 270]	KM
9	Maksymalna prędkość	przedział liczbowy	[150; 247]	km/h
10	Przyśpieszenie 0-100 km/h	przedział liczbowy	[6; 18]	s
11	Rodzaj paliwa	lista kategorii	{benzyna; diesel}	–
12	Zużycie paliwa	przedział liczbowy	[3,7; 9,0]	l/100 km

Źródło: opracowanie własne.

należność samochodu do segmentu podano zgodnie ze wskazaniem producentów. W wyniku agregacji danych uzyskano obiekty symboliczne II rzędu, opisane 12 zmiennymi symbolicznymi (tab. 2). Każdy obiekt obejmuje wszystkie dostępne dla wybranego modelu wersje nadwozia i silnika, np. Toyota Corolla, w zależności od wersji silnika, osiąga prędkość maksymalną w granicach 175-200 km/h, przy przyspieszeniu od 0 do 100 km/h rzędu 10,0-11,9 s.

W klasyfikacji zastosowano metody hierarchiczne (Warda i kompletnego połączenia) i optymalizacyjne (k -medoidów i DCLUST). Macierz odległości wyznaczono z wykorzystaniem miary odległości Ichino-Yaguchi (U_3). Ze względu na relatywnie niewielką liczebność zbioru obiektów (30) jako wartość graniczną przyjęto strukturę 8 klas. Rozważano wskazania czterech indeksów wyboru liczby klas, tj. Calińskiego i Harabasa (G1d), Bakera i Huberta (G2), *silhouette* (S) oraz Huberta i Levine (G3). Liczbę klas dla trzech pierwszych indeksów wskazują ich wartości maksymalne, natomiast dla ostatniego indeksu wartości minimalne. Dokonano porównania wskazań indeksów oraz oceny klas pod względem poziomu ich użyteczności. Sprawdzone także, w jakim stopniu odwzorowują one podział na cztery autosegmenty.

W pierwszym podejściu w zbiorze pozostawiono wszystkie zmienne. Brak eliminacji zmiennych zakłócających strukturę klas spowodował, że indeksy wskazują zróżnicowaną liczbę klas dla większości metod (zob. tab. 3, część a). Jedynie dla DCLUST wyniki są w miarę spójne. W tym przypadku indeksy G2 oraz S wskazały dwie klasy, natomiast indeksy G1d oraz G3 – sześć klas. Naturalną liczbę klas (4 klasy) wskazał jedynie indeks S (dla metody Warda) i indeks G3 (dla metody k -medoidów). Wartości indeksu S nie przekraczały jednak 0,37; co oznacza słabą strukturę klas [por. Walesiak 2004, s. 66].

W drugim podejściu przeprowadzono selekcję zmiennych z wykorzystaniem dwóch procedur formalnych. Najpierw zastosowano metodę grafową Ichino [1994], dedykowaną analizie danych symbolicznych [zob. też Wilk, Dudek 2009; Pełka, Wilk 2010]. Metoda pozostawiła zmienne 3, 4, 5, 6 i 12; odrzuciła natomiast zmienne związane między innymi z osiąganymi i parametrami technicznymi, stosowane w segmentacji samochodów. Poprawiła jednak zgodność wskazań indeksów (zob. tab. 3, część b). Indeksy G1d i G2 wskazały osiem klas dla metod hierarchicznych, a indeks G2 również dla k -medoidów, natomiast indeks G3 dla metody kompletnego połączenia. Choć sugerowana liczba klas nie jest zgodna z podziałem naturalnym, to klasy są dosyć dobrze separowane. Dwie spośród klas odpowiadają segmentom A i D, natomiast segment B został podzielony na dwie klasy, a segment C na trzy klasy. Niespójna jest jedynie klasa 8, w której znalazły się auta z segmentów B, C, i D.

Indeks S wskazał dwie klasy w większości metod (oprócz kompletnego połączenia), jednak nie przekroczył wartości 0,38; co oznacza słabą strukturę klas. Najlepszy podział uzyskano metodą k -medoidów, w którym klasa pierwsza skupia auta z segmentów A i B, a klasa druga reprezentuje auta segmentów C i D. Indeksy G1d i G2 wskazały natomiast sześć klas w metodzie DCLUST, jednak uzyskana klasyfikacja znacznie odbiega od podziału naturalnego.

Tabela 3. Wartości indeksów wyboru liczby klas

Metoda Indeks*	Warda			Kompletnego połączenia			k-medoidów			DCLUST						
	G1d	G2	G3	S	G1d	G2	G3	S	G1d	G2	G3	S				
a) podejście bez selekcji zmiennych																
2	42,58	0,776	0,450	0,362	43,26	0,484	0,401	0,279	60,92	0,731	0,437	0,348	69,77	0,724	0,427	0,346
3	76,68	0,772	0,358	0,267	55,30	0,669	0,412	0,257	64,11	0,747	0,406	0,271	52,38	0,620	0,414	0,187
4	59,16	0,809	0,557	0,285	59,16	0,809	0,557	0,285	57,50	0,652	0,404	0,162	59,59	0,592	0,387	0,155
5	77,63	0,864	0,473	0,254	44,84	0,810	0,586	0,257	64,47	0,775	0,516	0,200	53,66	0,623	0,397	0,191
6	85,60	0,831	0,451	0,255	67,55	0,883	0,496	0,247	82,07	0,818	0,461	0,241	75,23	0,714	0,347	0,188
7	76,58	0,852	0,483	0,240	81,48	0,864	0,494	0,236	94,84	0,840	0,420	0,212	58,71	0,665	0,355	0,131
8	101,17	0,908	0,406	0,235	79,28	0,899	0,478	0,242	90,19	0,868	0,395	0,210	48,79	0,647	0,362	0,105
b) podejście z selekcją zmiennych metodą grafową Ichino																
2	60,77	0,592	0,464	0,364	9,65	0,543	0,482	0,303	75,58	0,654	0,377	0,373	75,58	0,656	0,372	0,374
3	84,21	0,694	0,447	0,301	50,54	0,682	0,486	0,294	80,71	0,663	0,453	0,277	72,41	0,649	0,425	0,253
4	69,19	0,713	0,421	0,276	41,32	0,751	0,534	0,318	83,50	0,707	0,465	0,243	64,37	0,655	0,440	0,258
5	68,31	0,795	0,443	0,285	64,75	0,816	0,517	0,276	76,82	0,692	0,455	0,224	84,61	0,704	0,439	0,232
6	82,04	0,821	0,498	0,265	77,51	0,825	0,509	0,265	80,35	0,775	0,513	0,237	96,62	0,732	0,370	0,207
7	99,64	0,876	0,453	0,283	91,79	0,869	0,465	0,275	99,86	0,812	0,478	0,255	93,10	0,729	0,364	0,171
8	128,31	0,966	0,422	0,260	99,41	0,874	0,444	0,263	97,84	0,837	0,451	0,247	78,46	0,722	0,365	0,166
c) podejście z selekcją zmiennych metodą HINoV Carmono, Cara i Maxwell																
2	42,58	0,776	0,450	0,362	44,04	0,483	0,408	0,285	<i>60,92</i>	<i>0,731</i>	0,437	<i>0,348</i>	70,09	0,724	0,424	0,348
3	76,68	0,772	0,358	0,267	55,30	0,669	0,412	<i>0,257</i>	132,26	0,878	0,433	0,431	56,11	0,560	0,331	0,301
4	194,43	0,970	0,385	0,461	80,97	0,748	0,465	0,251	<i>57,50</i>	<i>0,652</i>	<i>0,404</i>	<i>0,162</i>	60,88	0,717	0,461	0,344
5	77,63	0,864	0,473	0,254	46,91	0,828	0,566	0,267	<i>64,47</i>	<i>0,775</i>	<i>0,516</i>	<i>0,200</i>	53,66	0,623	0,397	<i>0,191</i>
6	91,19	0,858	0,422	0,266	93,98	0,823	0,485	0,254	<i>82,07</i>	<i>0,818</i>	<i>0,461</i>	<i>0,241</i>	86,52	0,754	0,316	0,209
7	76,24	0,840	0,482	0,238	81,13	0,857	0,493	0,234	114,70	0,853	0,475	0,301	58,71	0,665	0,355	<i>0,131</i>
8	100,86	0,903	0,404	0,233	78,97	0,893	0,476	0,240	<i>90,19</i>	<i>0,868</i>	0,395	<i>0,210</i>	56,54	0,584	0,333	0,094

* adaptacja indeksu Calinskiego i Harabasz G1d, indeks Bakera i Huberta G2, indeks Huberta i Levine G3, indeks *silhouette* Rouseauwa S.

Objaśnienia: wartości wytuszone oznaczają optymalną wartość indeksu, wartości pisane kursywą w części c) oznaczają pełny zbiór zmiennych.

Źródło: opracowanie z wykorzystaniem pakietów clusterSim [Walesiak, Dudek 2013] i symbolica [Dudek i in. 2013] programu R.

Zastosowanie metody grafowej Ichino nie przyniosło zadowolających rezultatów. Z tego względu zastosowano alternatywną metodę selekcji zmiennych, tj. adaptację metody *HINoV* Carmone'a, Kary i Maxwell [1999] dla danych symbolicznych [zob. Walesiak, Dudek 2008], która wymaga zadania *a priori* metody klasyfikacji i liczby klas. Metoda, w większości przypadków, zachowała zmienne związane z osiąganiami i parametrami (tab. 4).

W zależności od metody klasyfikacji mierniki wskazywały zróżnicowaną liczbę klas (tab. 3, część 3). Indeksy G1d, G2 i S wskazały cztery klasy w podziale metodą Warda; indeks S uzyskał wartość bliską 0,5; co można interpretować jako poważną strukturę klas. Należy zauważyć, że w metodzie *k*-medoidów te same indeksy wskazały 3 klasy. Natomiast optymalne wartości indeksów G3 i S w klasyfikacji metodą kompletnego połączenia sugerują podział na 2 klasy, ale w DCLUST strukturę sześciu klas potwierdzają indeksy G1d, G2 i G3.

Tabela 4. Zmienne wybrane metodą *HINoV* Carmone'a, Kary i Maxwell

Liczba klas	Metody			
	Warda	Kompletnego połączenia	<i>k</i> -medoidów	DCLUST
2	wszystkie	oprócz 5	wszystkie	oprócz 11
3	wszystkie	wszystkie	1, 3, 4, 9	1, 3, 4, 8, 9
4	3, 4, 9	1, 3, 4, 12	wszystkie	3, 9, 10
5	wszystkie	oprócz 6	wszystkie	wszystkie
6	oprócz 6	oprócz 2, 6, 12	wszystkie	oprócz 6
7	oprócz 11	oprócz 11	3, 7, 8, 9	wszystkie
8	oprócz 11	oprócz 11	wszystkie	oprócz 6, 11

Źródło: opracowanie z wykorzystaniem pakietu `symbolicDA` [Dudek i in. 2013] programu **R**.

Podział zbioru obiektów na dwie klasy metodą kompletnego połączenia pozwolił uzyskać dosyć dobrze separowalne klasy. Do klasy pierwszej należą auta droższe, większe i wydajniejsze (głównie z segmentu C i D), natomiast do klasy drugiej auta tańsze, mniejsze i słabsze (segmenty A i B). Wyodrębnienie trzech klas metodą *k*-medoidów prowadziło natomiast do uzyskania niezbyt dobrze separowanych klas. W skupieniach znajdują się auta z różnych segmentów, np. w jednej klasie jest Octavia, Civic, Golf i Leon, znacznie różniące się gabarytami.

Podział na cztery klasy metodą Warda okazał się w znacznym stopniu zgodny z autosegmentami. Jednak według tej klasyfikacji w jednej klasie z Hondą Accord i Passatem znalazł się Leon i Golf ze względu na porównywalne osiągi. Najdokładniejszy, choć nie zgodny z naturalną liczbą klas, okazał się podział na sześć klas metodą DCLUST. W klasie pierwszej znalazły się auta segmentu A, a w klasie drugiej auta segmentu D. Segment B został podzielony na dwie klasy (auta droższe i wydajniejsze; auta tańsze i mniej wydajne). Dwie podgrupy wydzielono także wśród aut segmentu C (auta mniejsze, np. Golf, auta większe, np. Octavia).

5. Podsumowanie

Złożony charakter danych symbolicznych implikuje zastosowanie procedury klasyfikacji obejmującej uprzednie dokonanie selekcji zmiennych najlepiej dyskryminujących zbiorów obiektów oraz zastosowanie formalnych indeksów wyboru liczby klas. Indeksy dały zbliżone wskazania co do liczby klas po wyeliminowaniu zmiennych zakłócających metodą *HINoV*. Wskazania indeksów były jednak zróżnicowane w zależności od zastosowanej metody klasyfikacji. Decyzja co do ostatecznej liczby klas zależeć powinna od celu badania.

Literatura

- Anderberg M.R. (1973), *Cluster Analysis for Applications*, Academic Press Inc., New York.
- Bock H.H., Diday E. (red.) (2000), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin – Heidelberg.
- Carmone F.J., Kara A., Maxwell S. (1999), *HINoV: A new method to improve market segment definition by identifying noisy variables*, „Journal of Marketing Research”, November, vol. 36, s. 501-509.
- Diday E., Noirhomme-Fraiture M. (red.) (2008), *Symbolic data analysis and the Sodas software*, John Wiley & Sons, Chichester.
- Dudek A. (2007), *Cluster quality indexes for symbolic classification. An examination*, [w:] H.H.-J. Lenz, R. Decker (red.), *Advances in Data Analysis*, Springer, Berlin, s. 31-38.
- Everitt B.S., Landau S., Leese M. (2001), *Cluster Analysis*, Arnold, London.
- Gowda C.K., Diday E. (1994), *Symbolic clustering algorithm using similarity and dissimilarity measures*, [w:] E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy (red.), *New approaches in classification and data analysis*, Springer Verlag, Berlin – Heidelberg, s. 414-421.
- Grabiński T. (1992), *Metody taksonometrii*, Wyd. AE w Krakowie, Kraków.
- Hair J.F., Black W.C., Babin B.J., Anderson R.E., Tatham R.L. (2006), *Multivariate Data Analysis*, Pearson Prentice Hall, New Jersey.
- Hardy A., Lallemand P., *Determination of the number of clusters for symbolic objects described by interval variables*, [w:] K. Jajuga, A. Sokołowski, H.-H. Bock (red.), *Classification, clustering and data analysis*, Springer, Berlin – Heidelberg, s. 311-318.
- Hardy A. (2005), *Validation of unsupervised symbolic classification*, Proceedings of ASMDA 2005 Conference (asmda2005.enst-bretagne.fr/IMG/pdf/proceedings/379.pdf).
- Ichino M., *Feature selection for symbolic data classification*, [w:] E. Diday, Y. Lechevallier, P.B. Schader, B. Burtschy (red.), *New Approaches in Classification and data analysis*, Springer Verlag, Berlin – Heidelberg, s. 423-429.
- Jain A.K., Dubes R.C. (1998), *Algorithms for clustering data*, Prentice Hall, Englewood Cliffs, New Jersey.
- Lechevallier Y. (red.) (2001), *Scientific report for unsupervised classification, validation and cluster representation*, Analysis System of Symbolic official Data – Project number IST-2000-25161.
- Malerba D., Esposito F., Giovalle V., Tamma V. (2001), *Comparing Dissimilarity Measures for Symbolic Data Analysis*, [w:] P. Nanopoulos (red.), *New Techniques and Technologies for Statistics: Exchange of Technology and Know-how*, s. 473-481.
- Malerba D., Esposito F., Monopoli M. (2002), *Comparing dissimilarity measures for probabilistic symbolic objects*, [w:] A. Zanasi, C.A. Brebbia, N.F.F. Ebecken, P. Melli (red.), *Data Mining III*, „Series Management Information Systems”, vol. 6, WIT Press, Southampton, s. 31-40.
- Mali K., Mitra S. (2003), *Clustering and its validation in a symbolic framework*, Pattern Recognition Letters, 24, s. 2367-2376.

- Milligan G.W., *Clustering validation: results and implications for applied analyses*, [w:] P. Arabie, L.J. Hubert, G. de Soete (red.), *Clustering and classification*, World Scientific, Singapore 1996, s. 341-375.
- Milligan G.W., Cooper M.C. (1985), *An examination of procedures for determining the number of clusters in a data set*, *Psychometrika*, 50, s. 159-179.
- Pełka M., Wilk J., *Metody selekcji zmiennych symbolicznych w zagadnieniach klasyfikacji*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, *Taksonomia* 17, Wrocław 2010, 216-223.
- Punj G., Stewart D.W. (1983), *Cluster Analysis in Marketing Research: Review and Suggestions for Application*, „*Journal of Marketing Research*”, Mai, vol. 20, s. 134-148.
- Verde R., Lechevallier Y., Chavent M. (2003), *Symbolic clustering interpretation and visualization*, „*The Electronic Journal of Symbolic Data Analysis*”, vol. 1, no. 1.
- Walesiak M., Dudek A. (2008), *Identification of noisy variables for nonmetric and symbolic data in cluster analysis*, [w:] C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker (red.), *Data analysis, machine learning and applications*, Springer Verlag, Berlin – Heidelberg, s. 85-92.
- Walesiak M. (2004), *Problemy decyzyjne w procesie klasyfikacji zbioru obiektów*, [w:] J. Dziechciarz (red.), *Zastosowania metod ilościowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1010, *Ekonometria* 13, Wrocław, s. 52-71.
- Wedel M., Kamakura W.A. (1998), *Market Segmentation: Conceptual and Methodological Foundations*, Kluwer Academic Publisher, Dordrecht.
- Wilk J. (2010), *Cluster analysis methods in symbolic data analysis*, [w:] J. Pocięcha (red.), *Data Analysis Methods in Economic Research*, *Studia i Prace UE w Krakowie* nr 11, Kraków, s. 39-54.
- Wilk J., Dudek A. (2009), *Metody doboru zmiennych w procesie klasyfikacji obiektów symbolicznych*, [w:] J. Dziechciarz (red.), *Zastosowania metod ilościowych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 84, *Ekonometria* 27, Wrocław, s. 20-28.
- Wilk J. (2005), *Miary odległości obiektów opisanych zmiennymi symbolicznymi z wagami*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1126, „*Taksonomia* 13, Wrocław, s. 224-236.
- Wilk J. (2006), *Problemy klasyfikacji obiektów symbolicznych. Symboliczne miary odległości*, [w:] J. Garczarczyk (red.), *Ilościowe i jakościowe metody badania rynku. Pomiar i jego skuteczność*, *Zeszyty Naukowe AE* nr 71, Wydawnictwo AE w Poznaniu, Poznań, s. 69-83.

PROBLEM OF DETERMINING THE NUMBER OF CLUSTERS IN TAXONOMIC ANALYSIS OF SYMBOLIC DATA

Summary: The problem of selecting the number of clusters was examined in the paper. A procedure, which may support revealing the structure of objects set, was proposed. It was based on determining a range of the number of clusters, selecting the symbolic variables with the use of formal algorithms, applying hierarchical, as well as optimization methods of cluster analysis, and also statistical indices of selecting the number of clusters. Eliminating noisy variables with the use of *HINoV* method and then applying Ward's and DCLUST methods produced the best results. However, the recommendations of indices were diversified due to the method of clustering and its properties. A final decision of the number of clusters must be determined as regards the objective of research.

Keywords: symbolic data, cluster analysis, number of clusters, numerical taxonomy, symbolic taxonomy.