

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

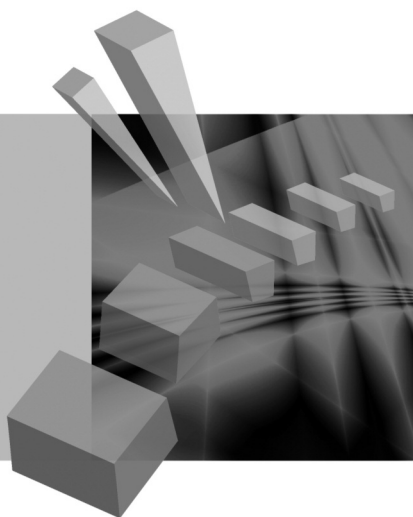
RESEARCH PAPERS

of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pocięcha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities <http://cejsh.icm.edu.pl>
oraz w 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](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

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomagania procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka , Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska , Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński , Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz , Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel , Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska , Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan , Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska , Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański , Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk , Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk , Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura , Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk , Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski , Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka , Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk , Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarc , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

Artur Mikulec

Uniwersytet Łódzki

METODY OCENY WYNIKU GRUPOWANIA W ANALIZIE SKUPIEŃ

Streszczenie: W artykule dokonano przeglądu metod oceny wyniku grupowania. Omówiono trzy metody wyboru właściwej liczby skupień dla metod aglomeracyjnych zaproponowane przez Mojenę i Wisharta oraz zaimplementowane w programie *ClustanGraphics 8*. Wspomniane kryteria bazują na relatywnych wartościach różnych poziomów połączeń obiektów na wykresie drzewa – *best cut significance test (upper tail, moving average)* oraz sprawdzaniu losowości podziału obiektów na wykresie drzewa – *tree validation*. Treść artykułu zilustrowana została przykładem empirycznym.

Słowa kluczowe: analiza skupień, ocena wyniku grupowania, kryteria Mojeny, kryterium Wisharta.

1. Wstęp

Metody oceny wyniku grupowania – w szerokim rozumieniu – związane są z trzema zagadnieniami analizy skupień: wyborem liczby skupień, porównywaniem wyników dwóch (i więcej) klasyfikacji oraz oceną jakości wyniku grupowania. W artykule dokonano zestawienia wybranych metod oceny wyniku grupowania, a przede wszystkim omówiono trzy metody wyboru właściwej liczby skupień dla metod hierarchicznych (aglomeracyjnych) zaproponowane przez Mojenę [1977] i Wisharta dostępne w programie *ClustanGraphics 8* [Wishart 2006]. Wspomniane kryteria bazują na: analizie odległości łączenia kolejnych obiektów na wykresie drzewa – *best cut significance test (upper tail rule, moving average quality control rule)* oraz ocenie losowości podziału obiektów na wykresie drzewa – *tree validation*. Przypomnienie i prezentacja tych metod wydają się zasadne z dwóch powodów: po pierwsze, metody aglomeracyjne są jednymi z najbardziej rozpowszechnionych i najczęściej wykorzystywanych w praktyce metod analizy skupień; po drugie, wyniki wielu analiz wskazują na subiektywne podejście autorów do wyboru liczby skupień na wykresie drzewa, tj. brak zastosowania formalnego kryterium dla tej oceny. Treść artykułu zilustrowana została przykładem empirycznym.

2. Metody oceny wyniku grupowania

Ogólny podział metod oceny wyniku grupowania stosowany w literaturze jest następstwem informacji wykorzystywanych w tego rodzaju analizie. Wyróżnia się metody sprawdzania wyniku grupowania oparte na kryterium wewnętrznym, zewnętrznym oraz względnym¹ [Gan i in. 2007]. Oceny na podstawie **kryterium wewnętrznego** dokonuje się tylko na bazie kryteriów ilościowych i informacji pochodzących z analizowanego zbioru danych. Dla pojedynczego wyniku klasyfikacji (indywidualnych skupień) uzyskanego metodami podziałowymi analizuje się stopień podobieństwa pomiędzy strukturą grupowania C , tj. przynależnością obiektów do skupień, a macierzą sąsiedztwa (odległości) obiektów P . W przypadku ciągu klasyfikacji uzyskanego metodami hierarchicznymi analizuje się ich strukturę, tzn. odległości łączenia kolejnych obiektów na wykresie drzewa. Porównuje się stopień podobieństwa macierzy odległości łączenia obiektów na wykresie drzewa H z macierzą sąsiedztwa (odległości) obiektów P – kryteria Mojeny. Inna metoda oceny wyniku grupowania hierarchicznego opiera się na testowaniu braku struktury klas [Gordon 1996] i może odbywać się na podstawie oceny losowości podziału obiektów na wykresie drzewa – kryterium Wisharta.

W podejściu według **kryterium zewnętrznego** ocenia się wyniki algorytmu grupowania uzyskane na podstawie wcześniej ustalonej struktury klas, która wynika z tego zbioru danych i odzwierciedla intuicyjnie jego strukturę. Ocena wyników grupowania odbywa się na podstawie struktury grupowania C i niezależnego podziału C_0 (wynik innego grupowania C_0) lub też na podstawie macierzy sąsiedztwa obiektów P i innego podziału C_0 (wynik innego grupowania C_0).

Uwzględniając trzy właściwe dla analizy skupień zagadnienia oceny wyniku grupowania, tj. poszukiwanie właściwej liczby skupień, porównywanie wyników dwóch klasyfikacji i ocenę jakości klasyfikacji, metody oceny wyniku grupowania można podzielić zgodnie ze schematem przedstawionym w tab. 1. Przegląd ten z pewnością nie wyczerpuje zbioru istniejących metod, zawiera jednak metody najczęściej stosowane, sprawdzone empirycznie oraz metody względnie nowe i stosunkowo słabo rozpowszechnione w polskiej literaturze i w praktyce badawczej. Tematem dalszej części artykułu będzie prezentacja trzech metod wyboru właściwej liczby skupień dla metod hierarchicznych (aglomeracyjnych) zaproponowanych przez Mojeny i Wisharta.

¹ Bez straty dla dalszych rozważań pominięto omawianie kryterium względnego, gdyż metody na nim oparte definiowane są w ten sam sposób jak metody oparte na kryterium wewnętrznym. Ocena wyniku grupowania w przypadku kryterium względnego opiera się na iteracyjnym wyborze najlepszego rezultatu grupowania ze zbioru pewnych dopuszczalnych rozwiązań.

Tabela 1. Metody oceny wyników analizy skupień^a

Wyszczególnienie	Kryterium	
	wewnętrznej informacji, oparte na ^b : { <i>C</i> , <i>P</i> }, { <i>H</i> , <i>P</i> }, braku struktury klas	zewnątrznej informacji, oparte na ^b : { <i>C</i> , <i>C</i> ₀ } lub { <i>P</i> , <i>C</i> ₀ }
Wybór właściwej liczby skupień	indeks: Bakera i Huberta; Beale'a (<i>F-ratio</i>); Calińskiego i Harabasza; Daviesa-Bouldina; Dudy i Harta; Dunna; <i>gap</i> (odstępu); Hartigana; Huberta i Levine'a; Krzanowskiego i Lai; <i>RMSSTD</i> ^c ; <i>RS</i> ; <i>SD</i> ; <i>S_Dbw</i> ; metoda: <i>jump</i> ; <i>ps</i> ; <i>clest</i> ; miara: Γ -Huberta; znormalizowana Γ -Huberta; kryterium: <u>Mojeny</u> – górnego obszaru odrzucenia (<i>upper tail rule</i>), średniej ruchomej (<i>moving average quality control rule</i>); <u>Wisharta</u> – losowości podziału obiektów na wykresie drzewa (<i>tree validation</i>); współczynnik: <u>korelacji kofenetycznej</u>	–
Porównywanie wyników dwóch klasyfikacji	<u>metryka Minkowskiego</u> ; <u>suma kwadratów odchyień</u> ; współczynnik: <u>korelacji kofenetycznej</u>	indeks: <u>Fowlkesa-Mallowsa</u> ; Jaccarda; Lermana; Wallace'a; miara: Γ -Huberta; znormalizowana Γ -Huberta; Randa; skorygowana Randa; wskaźnik: podobieństwa Nowaka
Ocena jakości klasyfikacji	indeks: sylwetkowy (<i>silhouette index</i>); miara: średniej zwartości skupień (<i>average of compactness</i>)	–

^a Przez podkreślenie wyróżniono metody służące do oceny wyniku grupowania hierarchicznego; ^b *C* – struktura grupowania, *P* – macierz sąsiedztwa (odległości) obiektów, *H* – macierz odległości łączenia obiektów na wykresie drzewa, *C*₀ – wynik innego grupowania obiektów; ^c Dla metod hierarchicznych należy stosować łącznie z indeksem *RS*.

Źródło: opracowanie własne na podstawie [Cormack 1971; Denoed i in. 2005; Fowlkes, Mallows 1983; Gan i in. 2007; Gatnar, Walesiak 2009; Gordon 1987; Kaufman, Rousseeuw 2005; Migdał-Najman, Najman 2008; Milligan, Cooper 1985; Mojena 1977; Nowak 1985; Sugar, James 2003; Tibshirani i in. 2001; Wishart 2006].

3. Kryterium *upper tail rule* oraz *moving average quality control rule*

Jedną z najbardziej znanych w literaturze przedmiotu prac poświęconych metodom wyboru liczby skupień jest artykuł [Milligan, Cooper 1985], dotyczący empirycznej analizy i oceny 30 procedur wyboru liczby skupień. Autorzy zidentyfikowali pięć najlepszych wówczas „reguł zatrzymania” (według kolejności): Calińskiego i Hara-

basza, Dudy i Harta, Huberta i Levine'a, Backera i Huberta oraz Beale'a (*F-ratio*)². W pierwszej dziesiątce omawianych procedur znalazło się też pierwsze kryterium Mojena – *upper tail rule* bazujące na względnych wartościach połączeń obiektów na wykresie drzewa.

Grupowanie N obiektów metodami hierarchicznymi (aglomeracyjnymi), które na początku stanowią N skupień, polega na krokowym łączeniu pojedynczych obiektów, tj. redukowaniu o 1 liczby skupień do czasu, aż wszystkie obiekty zostaną włączone do jednego skupienia. Wyniki grupowania hierarchicznego (aglomeracyjnego) N obiektów można opisać za pomocą zbioru kolejnych klasyfikacji P_0, P_1, \dots, P_{N-1} oraz korespondujących z nimi wartości kryterium klasyfikacji $\alpha_0, \alpha_1, \dots, \alpha_{N-1}$, gdzie krok (subskrypt) $j = 0, 1, \dots, N-1$ odpowiada uzyskanej kolejno $N, N-1, \dots, 1$ liczbie skupień. Struktura grupowania podziału P_j jest reprezentowana przez skupienia C_1, C_2, \dots, C_{N-j} .

Mojena zauważył, że różnice pomiędzy wynikami poszczególnych metod aglomeracyjnych są jedynie następstwem odmiennego sposobu definiowania miary połączenia klas (wszystkie metody opierają się na procedurze Lance'a i Williamsa i dla każdej z nich uzyskuje się inny ciąg klasyfikacji P_0, P_1, \dots, P_{N-1} i korespondujące z nimi wartości kryterium $\alpha_0, \alpha_1, \dots, \alpha_{N-1}$ – poziomu połączenia klas) oraz znaczenia, w jakim zdefiniowany jest parametr α . Jeżeli w analizie w charakterze miary połączenia przyjmie się miarę niepodobieństwa (odległości) między skupieniami, a więc jej większe wartości implikują większe niepodobieństwo grup, to [Mojena 1977]:

$$\alpha_j = \min_{i < m} [d_{im}], \quad i, m = 1, \dots, N - j, \quad (1)$$

w przeciwnym razie (przyjmując miarę podobieństwa między skupieniami):

$$\alpha_j = \max_{i < m} [d_{im}], \quad i, m = 1, \dots, N - j. \quad (2)$$

Statystyczne reguły służące selekcji najlepszego „podziału” badanej zbiorowości mogą być oparte na rozkładzie wartości kryterium α lub na odpowiedniej transformacji tego kryterium. Jeżeli większe wartości miary połączenia implikują niepodobieństwo jednostek, to rozkład wartości parametru α jest monotonicznie rosnący dla metod aglomeracyjnych: pojedynczego i kompletnego połączenia, średniej klasowej, ważonej średniej klasowej i Warda, przy zastosowaniu wzoru (1) – przypadek A oraz monotonicznie malejący dla tych metod aglomeracyjnych, przy zastosowaniu wzoru (2) – przypadek B. Stąd też Mojena zaproponował dwie metody, które bazują na ocenie „istotności” zmian wartości parametru α stanowiącego kryterium klasyfikacji.

Pierwsza reguła wykorzystuje $N-1$ wartości parametru α do wyznaczenia jego wartości średniej i odchylenia standardowego z próby celem zdefiniowania

² Kryteria Dudy i Harta oraz Beale'a są odpowiednie dla zagnieżdżonej struktury klas.

„istotnej” wartości α , która jako pierwsza leży w „górnym” – wzór (1) lub w „dolnym” – wzór (2) obszarze odrzucenia rozkładu jego wartości. Dla metod aglomeracyjnych: pojedynczego i kompletnego połączenia, średniej klasowej, ważonej średniej klasowej i Warda – wzór (1), przypadek A – ze zbioru kolejnych klasyfikacji P_1, \dots, P_{N-1} należy wybrać taką klasyfikację P_j , aby odpowiadający jej krok j ($j=1 \dots, N-2$) pierwszy spełniał nierówność sformułowaną dla poziomu połączenia klas α [Mojena 1977]:

$$\alpha_{j+1} > \bar{\alpha} + k \cdot s_{\alpha}, \quad (3)$$

natomiast dla przypadku B uwzględniającego wzór (2) sens nierówności (3) jest odwrotny – należy wybrać taką klasyfikację P_j , aby odpowiadający jej krok j , ($j=1 \dots, N-2$) pierwszy spełniał nierówność [Balicki 2009]:

$$\alpha_{j+1} < \bar{\alpha} - k \cdot s_{\alpha}, \quad (4)$$

gdzie α_{j+1} to poziom połączenia klas w kroku $j+1$, $\bar{\alpha}$ i s_{α} to odpowiednio średnia i nieobciążone odchylenie standardowe poziomu połączenia klas, a k to pewna stała³.

Jeżeli żadna z wartości α_j ($j=1, \dots, N-2$) poziomu połączenia klas nie spełnia nierówności (3) lub (4), to należy: (a) przyjąć, że wszystkie analizowane obiekty tworzą jedno skupienie – brak rozwiązania, (b) wybrać taki krok j , dla którego w kroku $j+1$ wartość $(\alpha_{j+1} - \bar{\alpha})/s_{\alpha}$ jest maksymalna, (c) wybrać inną regułę podziału obiektów⁴.

Druga reguła podziału zaproponowana przez Mojenę jest bardziej skomplikowana i bazuje na metodzie średniej ruchomej, mającej szerokie zastosowanie w kontroli jakości. Mojena zauważył, że zachowanie się wartości α (poziomu połączenia klas) nie jest odmienne od zachowania się wartości szeregu czasowego z trendem, co zasugerowało, że lepszym podejściem jest sformułowanie 1-krokowego modelu przewidującego, który zdefiniuje istotną wartość parametru α

³ Autor sugerował przyjąć w obliczeniach wartość $k \in (2,75; 3,50)$, z kolei Milligan i Cooper [1985] na podstawie przeprowadzonych badań empirycznych sugerowali wartość $k = 1,25$. Alternatywnie zamiast k można przyjąć wartość z rozkładu t -Studenta, co oznacza dodatkowe założenie, że α , tj. poziom połączenia klas na wykresie drzewa, podlega rozkładowi normalnemu.

⁴ Dla tej reguły Mojeny wybranie wariantu (a) lub (b) daje ten sam wynik dla monotonicznych wartości α_j , jeżeli $\bar{\alpha}$ i s_{α} są deterministyczne (ustalone), oraz różny wynik, jeżeli $\bar{\alpha}$ i s_{α} są stochastyczne.

w kroku $j+1$ ⁵. Dla metod aglomeracyjnych: pojedynczego i kompletnego połączenia, średniej klasowej, ważonej średniej klasowej i Warda – wzór (1), przypadek A – celem reguły jest wybór takiej klasyfikacji P_j , aby odpowiadający jej krok j ($j = r, r+1, \dots, N-2$) pierwszy spełniał nierówność [Mojena 1977]:

$$\alpha_{j+1} > \bar{\alpha}_j + L_j + b_j + k \cdot s_j, \quad (5)$$

gdzie: r oznacza liczbę wartości poziomu połączenia klas α w danym kroku – liczbę wartości do wyznaczania średniej ruchomej⁶, $\bar{\alpha}_j$ jest średnią ruchomą wartości parametru α obliczoną w kroku j , L_j jest korektą dla opóźnionego trendu poziomu połączenia klas obliczoną w kroku j , b_j jest „ruchomym” średniokwadratowym nachyleniem linii trendu poziomu połączenia klas w kroku j , k jest pewną stałą, a s_j jest „ruchomym” nieobciążonym oszacowaniem odchylenia standardowego wartości parametru α , przy czym wartości L_j i b_j wyraża się następująco:

$$L_j = \frac{(r-1)b_j}{2}, \quad (6)$$

$$b_j = \frac{6 \left[2 \sum_{f=j-r+1}^j w_f \alpha_f - (r+1) \sum_{f=j-r+1}^j \alpha_f \right]}{r(r^2-1)}, \quad (7)$$

gdzie: $w_f = w_{f-1} + 1$, $f = (j-r+2), \dots, j$ oraz $w_{j-r+1} = 1$.

Dla przypadku B opisanego wzorem (2) sens nierówności (5) jest odwrotny, należy wybrać taką klasyfikację P_j , aby odpowiadający jej krok j ($j = r, r+1, \dots, N-2$) pierwszy spełniał nierówność:

$$\alpha_{j+1} < \bar{\alpha}_j - L_j - b_j - k \cdot s_j. \quad (8)$$

⁵ Idea drugiej reguły polega na wyznaczaniu oczekiwanej wartości poziomu połączenia klas α na każdym j -tym etapie, na podstawie liniowej funkcji trendu dopasowywanej (korygowanej) względem pierwszych $j-1$ wartości poziomu połączenia klas na wykresie drzewa.

⁶ W regule drugiej również oblicza się średnią arytmetyczną i odchylenie standardowe wartości parametru α (poziomu połączenia klas) oznaczone odpowiednio $\bar{\alpha}_j$ i s_j , jednakże wyznacza się je sekwencyjnie po każdym kolejnym j -tym kroku procedury ($j = r, r+1, \dots, N-2$). Do wyznaczenia optymalnej liczby klas w regule pierwszej należy znać wszystkie $N-1$ wartości poziomu połączenia klas, zaś w regule drugiej wystarczy znać tylko r wartości poziomu połączenia klas ($r \leq N-2$) [Gatnar, Walesiak 2004].

Jeżeli żadna z wartości α_j ($j = r, r+1, \dots, N-2$) poziomu połączenia klas nie spełnia nierówności (5) lub (8), to procedura dalszego postępowania jest analogiczna jak w przypadku pierwszej reguły Mojeny – warianty (a), (b) i (c). Warto jednak zauważyć, że $\bar{\alpha}_j + L_j$ jest oczekiwaną wartością α_j , a $\bar{\alpha}_j + L_j + b_j$ jest oczekiwaną wartością α_{j+1} , stąd wartość znormalizowana obliczana w wariancie (b) równa się $(\alpha_{j+1} - \bar{\alpha}_j - L_j - b_j) / s_\alpha$.

4. Kryterium *tree validation*

Trzecie kryterium, którym warto zainteresować się w kontekście wyboru właściwej liczby skupień w przypadku metod hierarchicznych (aglomeracyjnych), opiera się na testowaniu braku struktury klas – ocenie losowości podziału obiektów na wykresie drzewa.

Idea metody *tree validation* (*bootstrap validation*) [Wishart 2006] polega na porównywaniu wyników ciągu klasyfikacji uzyskanych metodami aglomeracyjnymi, mających postać drzewa, z rodziną drzew generowanych na podstawie losowej permutacji tego samego zbioru danych albo skojarzonej macierzy sąsiedztwa. W tym celu wykorzystywane są techniki bootstrapowego, wielokrotnego losowania prób (bez zwracania), za pomocą których konstruowane są pojedyncze dendrogramy wyników grupowania [Domański i in. 1998]. Na każdym poziomie od 1 do N skupień, które są możliwe w przypadku grupowania N obiektów, oceniana jest jakość drzewa uzyskanego na podstawie analizowanych (oryginalnych) danych w stosunku do uśrednionej wersji drzewa wyznaczonej na podstawie prób losowych. Metoda oceny losowości podziału obiektów na wykresie drzewa stara się odrzucić hipotezę H_0 mówiącą o tym, że struktura grupowania obiektów w postaci danego drzewa jest losowa (brak struktury klas) na rzecz alternatywnej. Inaczej mówiąc, na podstawie analizowanego zbioru danych metoda *tree validation* poszukuje na wykresie drzewa, będącym ciągiem klasyfikacji, podziału, który jest „najbardziej oddalony” od podziału losowego.

5. Przykład empiryczny

Dla zilustrowania omówionych procedur wyboru właściwej liczby skupień (*upper tail rule*, *moving average quality control rule* i *tree validation*) programu *Clustan-Graphics 8* dokonano klasyfikacji przykładowego zbioru danych *Mammals.xls* (pochodzącego z płyty CD *ClustanGraphics 8*) zawierającego skład mleka 25 ssaków opisany za pomocą 5 cech, tj. zawartości: wody, białka, tłuszczu, laktozy oraz składników mineralnych (ash).

W pierwszym kroku wyznaczono macierz kwadratowych odległości euklidesowych. Następnie, wskazując metodę aglomeracyjną – Warda, uzyskano wyjściowy

wykres drzewa przedstawiający ciąg klasyfikacji. Po wybraniu z menu polecenia *Tree/Best Cut*, a w dalszej kolejności *Upper tail* otrzymano wszystkie „istotne” podziały dla aktualnego drzewa według pierwszego kryterium Mojeny (2 lub 3 skupienia). Na podstawie ciągu wartości poziomu połączenia klas obliczany jest średni poziom połączenia klas, jego odchylenie standardowe oraz statystyka *t*-Studenta (standaryzowane odchylenie poziomu połączeń klas od średniej). Z kolejnych poziomów połączenia klas wybierane jest pierwsze „istotne” na poziomie 5%, które jest automatycznie „przenoszone” na wykres drzewa.

Wybierając z menu polecenie *Tree/Best Cut*, a w dalszej kolejności *Moving average* otrzymano „istotne” podziały dla aktualnego drzewa według drugiego kryterium Mojeny, które w tym przypadku wskazało inny podział mleka ssaków – 9 skupień.

W ostatnim kroku, wybierając z menu polecenie *Tree/Validate* oraz decydując, czy drzewa mają być generowane na podstawie losowej permutacji zbioru danych czy macierzy sąsiedztwa, otrzymano wynik oceny losowości podziału obiektów na wykresie drzewa. Metoda *tree validation* wskazała, że „najbardziej oddalony” od podziału losowego jest podział mleka ssaków na 3 skupienia, jednorodne pod względem zawartości wody, białka, tłuszczu, laktozy oraz składników mineralnych (ash).

6. Podsumowanie

W artykule omówiono kryteria Mojeny i Wisharta wyboru właściwej liczby skupień dla metod hierarchicznych (aglomeracyjnych), a ich działanie zilustrowane zostało przykładem z programu *ClustanGraphics* 8. Należy zauważyć, że w załączonym przykładzie dwa spośród trzech kryteriów wskazały ten sam wynik grupowania – 3 skupienia, który wydaje się tym właściwym. Celem kolejnego artykułu z tego zakresu będzie empiryczna analiza i porównanie trzech omówionych kryteriów na tle innych powszechnie stosowanych miar wyboru optymalnej liczby skupień.

Literatura

- Balicki A., *Statystyczna analiza wielowymiarowa i jej zastosowania społeczno-ekonomiczne*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk 2009.
- Cormack R., *A review of classification*, „Journal of the Royal Statistical Society”, Series A 1971, vol. 134(3).
- Denoeud L., Garreta H., Guénoche A., *Comparison of distance indices between partition*, Conference „International Symposium on Applied Stochastic Models and Data Analysis” 2005 (ASMDA2005), <http://conferences.telecom-bretagne.eu/asmda2005>.
- Domański Cz., Pruska K., Wagner W., *Wnioskowanie statystyczne przy nieklasycznych założeniach*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 1998.
- Fowlkes E.B., Mallows C.L., *A method for comparing two hierarchical clusterings*, „Journal of the American Statistical Association” 1983, vol. 78(383).
- Gan G., Ma C., Wu J., *Data clustering: theory, algorithms, and applications*, SIAM, Philadelphia 2007.

- Gatnar E., Walesiak M. (red.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo AE, Wrocław 2004.
- Gatnar E., Walesiak M. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo PWN, Warszawa 2009.
- Gordon A.D., *A Review of hierarchical classification*, „Journal of the Royal Statistical Society”, Series A 1987, vol. 150(2).
- Gordon A.D., *Hierarchical Classification*, [w:] Arabie P., Hubert L.J., De Soete G. (red.), *Clustering and Classification*, World Scientific, Singapore 1996.
- Kaufman L., Rousseeuw P.J., *Finding Groups in Data. An Introduction to Cluster Analysis*, (reprint 2005), John Wiley & Sons, New York 2005.
- Migdał-Najman K., Najman K., *Wykorzystanie wskaźnika Dunna do ustalania optymalnej liczby skupień*, „Wiadomości Statystyczne” 2008, nr 11.
- Milligan G.W., Cooper M.C., *An examination of procedures for determining the number of clusters in a data set*, „Psychometrika” 1985, vol. 50(2).
- Mojena R., *Hierarchical grouping methods and stopping rules: an evaluation*, „Computer Journal” 1977, vol. 20(4).
- Nowak E., *Wskaźnik podobieństwa wyników podziałów*, „Przegląd Statystyczny” 1985, nr 1.
- Sugar C.A., James G.M., *Finding the number of clusters in a data set: an information-theoretic approach*, „Journal of the American Statistical Association” 2003, vol. 98(463).
- Tibshirani R., Walther G., Hastie T., *Estimating the number of clusters in a data set via the gap statistic*, „Journal of the Royal Statistical Society” 2001, Series B, vol. 63(2).
- Wishart D., *ClustanGraphics Primer: a Guide to Cluster Analysis*, (4th edition), Edinburgh 2006.

EVALUATION METHODS FOR THE GROUPING RESULT IN CLUSTER ANALYSIS

Summary: The paper reviews evaluation methods for the grouping result. It discusses three methods how to choose the correct number of clusters for agglomeration methods proposed by Mojena and Wishart, which are implemented in *ClustanGraphics 8*. The criteria which are mentioned are based on the relative sizes of the different fusion levels of the objects in the dendrogram – best cut significance test (*upper tail, moving average*) and checking randomness of objects clustering in the dendrogram – *validation tree*. The text is followed by an empirical example.

Keywords: cluster analysis, evaluation of the grouping result, Mojena criteria, Wishart criterion.