

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

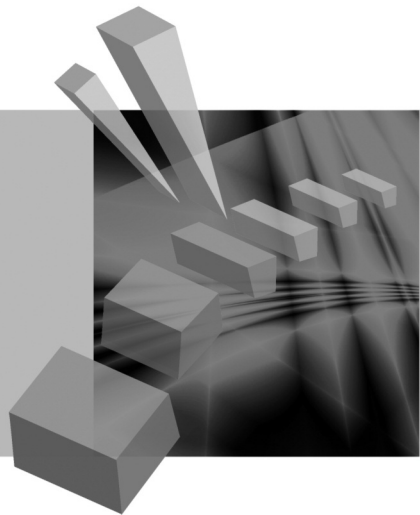
RESEARCH PAPERS

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,
Miroslaw 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
Sylvia 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 Szwarz , 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

Jerzy Korzeniewski

Uniwersytet Łódzki

OCENA EFEKTYWNOŚCI METODY UŚREDNIANIA ZMIENNYCH I METODY ICHINO SELEKCJI ZMIENNYCH W ANALIZIE SKUPIEŃ

Streszczenie: Metodę Ichino można stosować do zmiennych mierzonych na dowolnych skalach. Pelka i Wilk zbadali tę metodę, ale badanie ograniczało się do kilku modeli struktur skupień, zaś kryterium oceny była stabilność grupowania mierzona indeksem sylwetkowym. Metoda uśredniania zmiennych nieistotnych dla struktury skupień [Fraiman i in. 2008] jest nową metodą, niezbadaną dotychczas w żadnym obszerniejszym eksperymencie. W artykule przedstawione są wyniki badania symulacyjnego, w którym kryteriami są pamięć, precyzja i indeks Randa. Metody konkurencyjne, z którymi porównywana jest metoda Ichino, to HINoV oraz dwie modyfikacje HINoV (VSKM i VAF z indeksem skupialności).

Słowa kluczowe: analiza skupień, wybór zmiennych, graf wzajemnego sąsiedztwa, uśrednianie zmiennych.

1. Wstęp

W kontekście analizy skupień ważnym zagadnieniem jest wybór zmiennych istotnych dla struktury skupień zbioru danych. Jeśli wybór zostanie dokonany błędnie lub zbiór zmiennych nie zostanie w ogóle oczyszczony ze zmiennych nieistotnych, to fakt ten na ogół wpływa bardzo negatywnie na wnioski uzyskane z analizy skupień. Do chwili obecnej zaproponowano kilkanaście podejść do zagadnienia selekcji zmiennych. Metody charakteryzują się różnymi cechami. Niektóre uzależniają wybór od jakiejś metody grupowania obserwacji, inne nie, niektóre są skonstruowane z myślą o jednej skali pomiarowej, inne są bardziej ogólne, niektóre mają charakter modelowy, tj. taki, w którym zakładamy, że zbiór danych jest mieszaniną obserwacji z rozkładów normalnych, inne są podejściami czysto heurystycznymi. W roku 2008 Steinley i Brusco przeprowadzili obszerny eksperyment symulacyjny, w którym dokonali porównania ośmiu metod służących do wybierania zmiennych w analizie skupień. Wybrali oni kilka podejść modelowych, takich jak: metoda wyrazistości cech (*feature saliency method* [Law i in. 2003]); metoda oparta na wielkości rozproszenia (*scatter separability method* [Dy, Brodley 2000]); metoda oparta na wyborze właściwego modelu (*model selection method* [Raftery, Dean 2006]) oraz kilka

niemodelowych: metoda HINoV [Carmone i in. 1999], metoda oparta na grupowaniu obiektów na wybranych podzbiorach zmiennych (COSA [Friedman, Meulman 2004]); metoda kolejnych rzutowań (*projection pursuit* [Montanari, Lizzani 2001]); metoda oparta na grupowaniu k -średnich (VS-KM *method* [Brusco, Cradit 2001]); metoda oparta na grupowaniu k -średnich z indeksem skupialości (VAF, *relative clusterability weighting method* [Steinley, Brusco 2007]). Ocena efektywności oparta była na trzech głównych wskaźnikach: pamięci (*recall*), precyzji (*precision*) oraz asymptotycznej odzyskiwalności poprawnego przypisania obserwacji do skupień (ARI, *asymptotic recovery*). Wskaźniki te definiuje się na potrzeby eksperymentu symulacyjnego, w którym dla każdego zbioru znany jest zbiór zmiennych istotnych dla struktury skupień oraz zbiór zmiennych nieistotnych. Wówczas przez **pamięć** rozumiemy stosunek liczby wybranych zmiennych istotnych do liczby wszystkich zmiennych istotnych, **precyzja** to stosunek liczby wybranych zmiennych istotnych do liczby wszystkich wybranych zmiennych, zaś **asymptotyczna odzyskiwalność** rozumiana jest w sensie średniej arytmetycznej (ze wszystkich zbiorów) wartości skorygowanego indeksu Randa mierzącego zgodność podziału opartego na wybranym podzbiorze zmiennych z podziałem wynikającym ze sposobu generowania zbioru danych. Wartość indeksu Randa obliczana jest dla podziału zbioru obiektów otrzymanego w następujący sposób: za pomocą metody k -średnich z losowym wyborem obserwacji startowych powtarzamy 100 razy grupowanie i zapamiętujemy grupowanie o najmniejszej wariancji wewnątrzgrupowej.

Od czasu cytowanego badania Steinleya i Brusco pojawiły się nowe metody. Istnieją również inne podejścia, wcześniejsze, niezbadane w cytowanym eksperymencie.

W artykule tym przedstawione są wyniki badań efektywności metody Ichino [1994] oraz metody uśredniania zmiennych [Fraiman i in. 2008]. Metoda Ichino ma bardzo ogólny charakter względem rodzaju skal pomiarowych, na których mierzone są zmienne opisujące obiekty – można ją stosować do dowolnych skal. Pełka i Wilk zbadali tę metodę, ale badanie ograniczało się do kilku modeli struktur skupień tworzonych przez zmienne ciągłe. Kryteria oceny były inne od tych zastosowanych przez Steinleya i Brusco [2008]. Efektywność oceniana była w sensie stabilności grupowania ocenianej za pomocą indeksu sylwetkowego (por. [Pełka, Wilk 2010]). Metoda uśredniania zmiennych nieistotnych dla struktury skupień jest nową metodą niezbadaną dotychczas w żadnym obszerniejszym eksperymencie. Podstawą wyników badań przedstawionych w dalszym ciągu artykułu jest eksperyment symulacyjny podobny do eksperymentu Steinleya i Brusco [2008]. Zastosowano te same miary efektywności. Obie metody oceniane były na tle trzech metod, które w badaniu Steinleya i Brusco [2008] wypadły najlepiej, tj. HINoV oraz dwóch modyfikacji HINoV (metody VSKM i VAF z indeksem skupialności).

2. Charakterystyka metody Ichino

Ichino opracował bardzo ogólny model populacji – model przestrzeni kartezjańskiej (*cartesian space model*). Model ten może być stosowany do zbiorów danych z szerokim zakresem kategorii zmiennych, od ciągłych do nominalnych. Model oparty jest na pojęciu operatora sumy obiektów \oplus oraz operatora iloczynu obiektów \otimes . Modelem przestrzeni kartezjańskiej nazywamy trójkę (U^d, \oplus, \otimes) , gdzie U^d jest d -wymiarową przestrzenią wartości zmiennych losowych opisujących obiekty zbioru danych

$$U^d = U_1 \times U_2 \times \dots \times U_d, \quad (1)$$

natomiast operator sumy i iloczynu definiujemy w następujący sposób. Sumę kartezjańską dwóch obiektów A oraz B definiuje się w postaci iloczynu kartezjańskiego operatora sumy poszczególnych współrzędnych:

$$A \oplus B = (A_1 \oplus B_1) \times (A_2 \oplus B_2) \times \dots \times (A_d \oplus B_d). \quad (2)$$

Operator iloczynu dwóch obiektów A oraz B definiujemy w postaci iloczynu kartezjańskiego operatora iloczynu poszczególnych współrzędnych:

$$A \otimes B = (A_1 \otimes B_1) \times (A_2 \otimes B_2) \times \dots \times (A_d \otimes B_d). \quad (3)$$

W dalszym ciągu definiujemy obiekty wzajemnie sąsiadujące (*mutual neighbours*) A_1 oraz A_2 względem zbioru obiektów symbolicznych $B = \{B_1, B_2, \dots, B_m\}$ jako obiekty spełniające warunek:

$$\forall B_i \in B, \quad B_i \otimes (A_1 \oplus A_2) = \emptyset. \quad (4)$$

Zbiór obiektów symbolicznych $A = \{A_1, A_2, \dots, A_k\}$ nazywamy grafem wzajemnego sąsiedztwa (*mutual neighbourhood graph*) względem zbioru obiektów symbolicznych $B = \{B_1, B_2, \dots, B_m\}$, gdy każda para obiektów ze zbioru A jest wzajemnie sąsiadująca względem zbioru B . Grafy wzajemnego sąsiedztwa pozwalają opisywać separowalne skupienia obiektów. Ichino, opierając się na takim sposobie badania podobieństwa grup obiektów, zaproponował kilka metod selekcji oraz grupowania zmiennych.

W przeprowadzonym dalej eksperymencie zastosowano następującą metodę.

1. Dla każdego podzbioru s zmiennych grupujemy wszystkie obiekty zbioru danych w 2 skupienia za pomocą metody k -średnich z losowym wyborem obiektów startowych (losujemy 50 razy i zapamiętujemy grupowanie z najmniejszą wariancją wewnątrzgrupową).

2. Dla każdego s znajdujemy największą liczbę par $l(s)$ wzajemnie sąsiadujących względem drugiego skupienia.

3. Za najlepiej odzwierciedlający strukturę skupień uznajemy ten podzbiór s zmiennych, dla którego różnica $l(s) - l(s-1)$ ma wartość najwyższą.

Pełka i Wilk sformułowali nieco inaczej metodę Ichino. Dla każdego podzbioru zmiennych konstruowany był zbiór grafów wzajemnego sąsiedztwa, po czym wystarczyło zliczyć liczbę par w każdym grafie (nie musimy sprawdzać, czy pary są wzajemnie sąsiadujące, gdyż zapewnia to algorytm szukania grafów).

3. Charakterystyka metody uśredniania zmiennych

Podstawą metody jest spostrzeżenie, że zmienne nieistotne dla struktury skupień powinny mieć we wszystkich skupieniach takie same rozkłady i, co za tym idzie, wartości średnie. Wobec tego, gdy zbiór danych został podzielony na pewną liczbę skupień, wystarczy sprawdzić, jak często podział oparty na wszystkich zmiennych pokrywał się z podziałem opartym na wybranym podzbiornie zmiennych, który uznamy za zbiór zmiennych istotnych, przy wartościach pozostałych zmiennych równych wartościom średnim.

Szczegółowe sformułowanie algorytmu metody jest następujące.

1. Dla zbioru wszystkich di zmiennych oryginalnych grupujemy wszystkie obiekty zbioru danych w tę samą liczbę skupień za pomocą metody k -średnich z losowym wyborem obiektów startowych (losujemy 50 razy i zapamiętujemy grupowanie z najmniejszą wariancją wewnątrzgrupową).

2. Dla każdego podzbioru s zmiennych ($s < di$) zmodyfikowanych (s zmiennych ma wartości oryginalne, zaś wartości pozostałych $di-s$ zmiennych zastępujemy wartościami średnimi) przypisujemy każdy obiekt do najbliższego centrum skupienia wyznaczonego w kroku pierwszym.

3. Spośród wszystkich podzbiorów s zmiennych za najlepiej odzwierciedlający strukturę skupień uznajemy minimalny podzbiór, który ma największą zgodność grupowania z grupowaniem uzyskanym dla zbioru wszystkich zmiennych.

Zgodność grupowania mierzymy za pomocą zmodyfikowanego indeksu Randa.

4. Eksperyment badawczy

W celu zachowania porównywalności badania przeprowadzono eksperyment na wzór eksperymentu Steinleya i Brusco [2008]. Poszerzono nieco zakres eksperymentu, dopuszczając zbiory z 2 i 3 skupieniami i maskujące rozkłady równomierne. Wszystkie zbiory składały się z 200 obiektów, różniły się między sobą następującymi cechami.

- Pierwsza cecha: liczba skupień, może być równa – 2, 3, 4, 6 lub 8.
- Druga cecha: liczebności skupień, możliwe są trzy warianty: (a) równe liczebności wszystkich skupień; (b) 10% obserwacji i (c) 60% obserwacji w jednym skupieniu, a pozostałe skupienia równoliczne.
- Trzecia cecha: liczba zmiennych istotnych, może być równa 2, 4 lub 6.
- Czwarta cecha: prawdopodobieństwo „zachodzenia na siebie” (*overlap*) skupień na każdej ze zmiennych istotnych, może być równe – 0, 0.1, 0.2, 0.3, 0.4. Se-

parowalność skupień jest typu “łańcuchowego”, tj. na każdym wymiarze jest $k - 1$ par skupień (k – liczba skupień), przy czym każde dwa kolejne zachodzą na siebie w stopniu (jednakowym dla wszystkich par), na który wskazuje prawdopodobieństwo.

- Piąta cecha: siła korelacji wewnątrz skupień, możliwe są dwa warianty: (a) macierz kowariancji w każdym skupieniu jest macierzą jednostkową; (b) w każdym skupieniu jest taka sama macierz kowariancji z jedynkami na przekątnej, zaś poza przekątną jest liczba wylosowana z odcinka [0.3; 0.8].
- Szósta cecha: liczba zmiennych maskujących, może być równa – 2, 4 lub 6.
- Siódma cecha: rozkład zmiennych maskujących. Możliwych jest siedem wariantów: (a) wszystkie zmienne niezależnie wygenerowane z rozkładu skośnego jednomodalnego (rozkład gamma z jednym stopniem swobody dla licznika i jednym dla mianownika); (b) wszystkie zmienne niezależnie wygenerowane z rozkładu normalnego ze średnią zero i jednostkową wariancją; (c) wszystkie zmienne niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0.25 poza przekątną; (d) wszystkie zmienne niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0.5 poza przekątną; (e) wszystkie zmienne niezależnie wygenerowane z rozkładu normalnego ze średnim wektorem zerowym i jedynkami na przekątnej w macierzy kowariancji i 0.75 poza przekątną; (f) wszystkie zmienne wygenerowane z niezależnych rozkładów normalnych ze średnimi równymi zero i wariancjami losowanymi z odcinka [1; 20]; (g) wszystkie zmienne niezależnie wygenerowane z rozkładów równomiernych na odcinku [1; 20]. Po uwzględnieniu wszystkich układów parametrów otrzymujemy razem liczbę 11 550 zbiorów.

5. Wyniki i wnioski

Wyniki przedstawione w tab. 1 pozwalają stwierdzić, że obie metody okazały się dużo gorsze od metod HINoV, VSKM oraz VAF. Te trzy metody, na tych samych zbiorach danych, uzyskały pamięć i precyzję rzędu 0,8-0,9 oraz wskaźnik *ARI* rzędu 0,7. Dla metody Ichino oraz metody uśredniania zmiennych wartości wskaźników są dużo gorsze. W przypadku metody Ichino można uznać, że wyniku takiego można było się spodziewać, gdyż autor tej metody raczej konstruował ją z myślą o słabszych skalach pomiarowych. We wszystkich publikacjach Ichino dominują przykłady ze zbiorami, w których występują tylko zmienne mierzone na skali nominalnej. Metoda jest bardzo czasochłonna, konieczne jest zbadanie wszystkich możliwych podzbiorów zbioru wszystkich zmiennych.

Słaba efektywność metody uśredniania zmiennych jest zaskakująca, gdyż zdaniem autorów metoda jest dobra. Porównując badanie autorów metody z naszym eksperymentem, należy zauważyć, że rozważali oni jedynie prosty przypadek dwóch zmiennych tworzących trzy raczej bardzo wyraźnie separowane skupienia. Odpo-

Tabela 1. Średnia pamięć, precyzja i asymptotyczna odzyskiwalność dla obu metod w zależności od liczby zmiennych nieistotnych dla struktury skupień oraz występowania w skupieniach korelacji wewnątrzgrupowej

Liczba zmiennych nieistotnych	Występowanie korelacji wewnątrz skupień	Metoda Ichino			Metoda uśredniania zmiennych		
		pamięć	precyzja	ARI	pamięć	precyzja	ARI
2	Nie	0,947	0,696	0,508	0,787	0,688	0,481
4	Nie	0,951	0,543	0,429	0,824	0,531	0,408
6	Nie	0,950	0,448	0,407	0,838	0,433	0,366
2	Tak	0,946	0,698	0,499	0,799	0,701	0,439
4	Tak	0,949	0,543	0,426	0,846	0,547	0,384
6	Tak	0,957	0,437	0,400	0,847	0,451	0,350
Średnio		0,933	0,561	0,499	0,821	0,559	0,405

Źródło: obliczenia własne.

wiadałoby to niemalże zbiorom z „overlap” równym 0 w naszym badaniu. Dla tych zbiorów wskaźnik ARI jest znacznie wyższy, równy ok. 0,7. Bardzo duży jest wtedy również odsetek zbiorów z bezbłędnym grupowaniem. Wadą metody jest jednak to, że niekiedy mamy do wyboru kilka podzbiorów zmiennych mających dokładnie taką samą wartość kryterium selekcji. Wówczas w algorytmie przyjęto, że wybrać należy podzbiór najmniej liczny. Ten sposób wybierania jednego z podzbiorów bardzo negatywnie wpływa na średnią precyzję, jaką uzyskała metoda. Jeżeli w przypadku równej wartości kryterium selekcji wybieralibyśmy wszystkie zmienne wchodzące w skład podzbiorów dających taką samą wartość kryterium, to precyzja metody byłaby nieco wyższa. Ciekawe jest to, że w przypadku występowania korelacji wewnątrz skupień metoda spisuje się wyraźnie gorzej. Zapewne gdy skupienia są „wydłużone”, to przypisywanie obiektu do wyznaczonego wcześniej (przy uwzględnieniu wszystkich zmiennych) centrum skupienia jest częściej błędne. Metoda w rozważanej postaci jest czasochłonna, ale dla dużych zbiorów zmiennych autorzy proponują algorytm filtrujący zbiór wszystkich zmiennych, typu *forward-backward search*, pracujący na zasadzie dołączania pojedynczych zmiennych do zbioru istniejącego. Ten algorytm został jednak zbadany tylko na dwóch empirycznych zbiorach danych. Efektywność obu metod bardzo szybko spada wraz ze wzrostem liczby zmiennych nieistotnych dla struktury skupień. Należy nadmienić, że dla metody uśredniania zmiennych badanie przeprowadzone było przy założeniu znajomości liczby skupień. To założenie w praktyce zbiorów empirycznych na ogół nie jest spełnione.

Literatura

- Brusco M., Cradit D., *A variable-selection heuristic for k-means clustering*, "Psychometrika" 2001, no 66.
- Carmone F.J.Jr., Kara A., Maxwell S., *HINoV: a new model to improve market segment definition by identifying noisy variables*, „Journal of Marketing Research” 1999, vol. 36.
- Dy J., Brodley C., *Feature subset selection and order identification for unsupervised learning*, Proc. 17th International Conf. on Machine Learning, 2000.
- Fraiman R., Justel A., Svarc M., *Selection of variables for cluster analysis and classification rules*, „JASA” 2008, no 103.
- Friedman J., Meulman J., *Clustering objects on subsets of attributes*, „Journal of the Royal Statistical Society” 2004, Series B 66.
- Ichino M., *Feature Selection for Symbolic Data Classification*, [w:] *New Approaches in Classification and Data Analysis*, Springer-Verlag, 1994.
- Law M., Jain A., Figueiredo M., *Feature Selection in Mixture-Based Clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003.
- Montanari A., Lizzani L., *A projection pursuit approach to variable selection*, „Computational Statistics and Data Analysis” 2001, vol. 35(4).
- Pełka M., Wilk J., *Metody selekcji zmiennych symbolicznych w zagadnieniach klasyfikacji*, [w:] *Taksonomia 17, Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, UE, Wrocław 2010.
- Raftery A.E., Dean N., *Variable selection for model based clustering*, „JASA” 2006, no 101.
- Steinley D., Brusco M., *A new variable weighting and selection procedure for k-means cluster analysis*, „Psychometrika” 2007.
- Steinley D., Brusco M., *Selection of variables in cluster analysis: an empirical comparison of eight procedures*, „Psychometrika” 2008, no 73.

EFFICIENCY ASSESSMENT OF ICHINO METHOD AND MEAN VALUE METHOD OF SELECTING VARIABLES IN CLUSTER ANALYSIS

Summary: Ichino method can be applied to variables measured on any type of scale. Pełka and Wilk examined the method but their investigation was limited to a couple of cluster structures and the criterion used was the stability of grouping based on the silhouette index. The mean value method [Fraiman et al. 2008] is a relatively new method not investigated until now in a broad simulation experiment. In the paper the results of simulation experiment are presented. The criteria used are: recall, precision and Rand index. The competing methods are HINoV and its two modifications: VSKM and VAF with clusterability index.

Keywords: cluster analysis, variable selection, mutual neighbourhood graph, taking mean values of variables.