

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

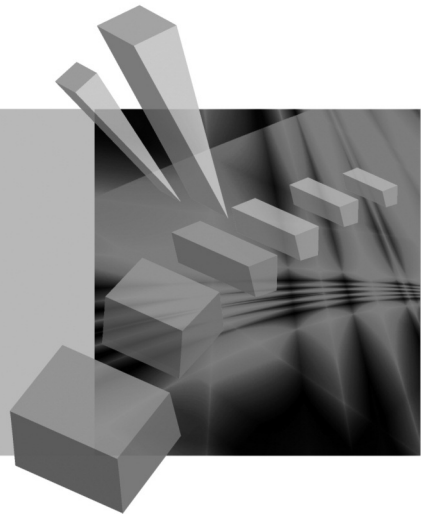
RESEARCH PAPERS

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 wspomaganie 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 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

Justyna Wilk

Uniwersytet Ekonomiczny we Wrocławiu

ANALIZA PORÓWNAWCZA OPROGRAMOWANIA KOMPUTEROWEGO W KLASYFIKACJI DANYCH SYMBOLICZNYCH

Streszczenie: Celem artykułu jest ocena przydatności dostępnego na rynku oprogramowania statystycznego w klasyfikacji danych symbolicznych. W artykule wyjaśniono podstawowe pojęcia analizy danych symbolicznych i omówiono procedurę klasyfikacji danych symbolicznych. Scharakteryzowano oprogramowanie statystyczne, jakie można stosować w klasyfikacji danych symbolicznych i wskazano jego użyteczność w poszczególnych jej etapach. Ze względu na relatywnie duże możliwości aplikacyjne bliżej scharakteryzowano pakiety R i SODAS.

Słowa kluczowe: klasyfikacja, dane symboliczne, oprogramowanie.

1. Wstęp

Analiza danych symbolicznych (ADS) jest relatywnie młodym kierunkiem rozwoju metod analizy eksploracyjnej i reprezentacji danych empirycznych. Złożoność zjawisk sprawia, że dane reprezentowane przez pojedynczą kategorię lub liczbę rzeczywistą stają się niewystarczające w opisie rzeczywistości. W ADS dopuszcza się występowanie danych w postaci przedziałów liczbowych, zbiorów kategorii i struktur udziałowych. Z tego względu metodologia ADS znajduje zastosowanie w badaniach dotyczących różnorodnych zagadnień (ekonomicznych, chemicznych, geologicznych itd.).

Przedmiotem wielu badań (m.in. marketingowych, finansowych, regionalnych) jest klasyfikacja obiektów (np. konsumentów, regionów) na podstawie zestawu kryteriów. Jednymi z najbardziej uznanych i efektywnych narzędzi klasyfikacji są metody klasyfikacji.

Użyteczność metodologii klasyfikacji danych symbolicznych w badaniach empirycznych jest w znacznej mierze podyktowana dostępnością oprogramowania komputerowego. Ze względu na złożoność tego rodzaju danych i wieloetapowość procedury klasyfikacji w analizie zastosowanie znajdują wybrane pakiety statystyczne. W literaturze, zarówno polskiej, jak i zagranicznej, brakuje opracowania na ten temat.

Celem artykułu jest dokonanie przeglądu dostępnego na rynku oprogramowania statystycznego i ocena jego przydatności w klasyfikacji danych symbolicznych. W artykule wyjaśniono podstawowe pojęcia ADS i omówiono procedurę klasyfikacji danych symbolicznych. Scharakteryzowano oprogramowanie statystyczne, jakie można stosować w klasyfikacji danych symbolicznych i wskazano jego użyteczność w poszczególnych jej etapach.

2. Specyfika danych symbolicznych

W ADS wyróżnia się obiekty i zmienne symboliczne. Obiektem symbolicznym określa się obiekt opisany zmiennymi symbolicznymi. Realizacjami zmiennej symbolicznej mogą być (por. [Bock, Diday 2000]):

- przedziały liczbowe (*interval-valued variable*) – zbiory wartości ciągłych ze zbioru liczb rzeczywistych, o równych bądź różnych rozpiętościach, np. przedziały wiekowe i dochodowe respondentów, przedział cenowy poszukiwanego produktu,
- zbiory kategorii (*multivalued variable*) – zestawy kategorii (równorzędnych lub uporządkowanych), wartości skokowych bądź przedziałów liczbowych, np. znajomość języków obcych, posiadane kategorie prawa jazdy, dostępne w sklepie rozmiary obuwia,
- struktury udziałowe (*modal variable*) – zbiory kategorii z przypisanymi indeksami wagowymi, prawdopodobieństwami, częstościami i udziałami procentowymi, np. struktura portfela inwestycyjnego, struktura miesięcznych wydatków konsumenta.

Uwzględnia się również występowanie logicznych powiązań między zmiennymi, w tym taksonomicznych (*taxonomic dependent variable*), hierarchicznych (*mother-daughter variable*) i logicznych (*logical dependent variable*).

Obiekty symboliczne, ze względu na stopień agregacji danych, można podzielić na (por. [Bock, Diday 2000]):

- obiekty w ujęciu klasycznym, tj. elementarne jednostki badania (*first order symbolic objects*), np. takie jak respondent, województwo, produkt,
- obiekty złożone (*second order symbolic objects*), będące wynikiem agregacji zbioru obiektów w ujęciu klasycznym.

Zbiór realizacji zmiennych symbolicznych dla obiektów zapisuje się w tablicy danych symbolicznych.

3. Klasyfikacja danych symbolicznych

Klasyfikacja jest złożonym procesem, którego wyniki zależą od wyborów dokonanych w każdym etapie. Typowa procedura klasyfikacyjna obejmuje następujące kroki (por. [Walesiak 2004; Punj, Stewart 1983, s. 144]):

1. Wybór obiektów i zmiennych.

2. Pomiar odległości i grupowanie obiektów.
3. Określenie liczby klas i ocena wyników klasyfikacji.
4. Interpretacja i profilowanie klas.

Początkowym etapem procedury klasyfikacyjnej jest ustalenie, w zależności od celu badania, jednostki badawczej oraz struktury i liczebności próby. W ADS za jednostkę badania można przyjąć obiekt w ujęciu klasycznym lub obiekt złożony. Następnie określa się zestaw zmiennych, na podstawie których przeprowadzona zostanie klasyfikacja. Dobór zmiennych ma przede wszystkim charakter merytoryczny. Niekiedy w selekcji zmiennych stosuje się również algorytmy formalne – dla zmiennych symbolicznych, m.in. metodę grafową Ichino czy adaptację metody *HINoV* Carmone, Kara i Maxwell.

Pomiar odległości jest uzasadniony w sytuacji, kiedy stosuje się metody bazujące na macierzy odległości, a wybór miary odległości zależy od charakteru zbioru zmiennych. W pomiarze odległości obiektów symbolicznych zastosowanie mają m.in. miary Ichino-Yaguchiego i de Carvalho.

Grupowanie obiektów przeprowadza się z wykorzystaniem metod taksonomicznych. Wielość i różnorodność procedur klasyfikacyjnych powoduje, że nie można wskazać metody uniwersalnej dla wszystkich problemów badawczych. Wybór metody jest determinowany celem badania. Grupowanie może mieć charakter hierarchiczny lub niehierarchiczny. Procedury hierarchiczne dają w wyniku hierarchię klas, którą uzyskuje się w drodze aglomeracji bądź deglomeracji zbioru obiektów. Z kolei metody niehierarchiczne (m.in. metody optymalizacyjne) dokonują podziału zbioru obiektów ze względu na przyjęte kryterium jakości podziału. W ADS zastosowanie mają (por. [Wilk 2010]):

- metody taksonomii numerycznej bazujące na macierzy odległości (zob. [Anderberg 1973; Grabiński, Wydymus, Zeliaś 1989; Everitt, Landau, Leese 2001]), szczególnie procedury hierarchiczne oraz niektóre metody optymalizacyjne,
- metody taksonomii symbolicznej bazujące na tablicy danych symbolicznych lub macierzy odległości (zob. [Gatnar 1998; Bock, Diday 2000; Verde 2004; Diday, Noirhomme-Fraiture 2008]).

Wybór liczby klas jest podyktowany wiedzą merytoryczną bądź wsparty metodami formalnymi. W ADS zastosowanie mają indeksy bazujące na macierzy danych i medoidach (np. indeks Calińskiego i Harabasz), macierzy odległości (np. indeks Bakera i Huberta) oraz tablicy danych symbolicznych (np. indeks $Q(P)$ Verde, Lechevallier i Chavent). Uzyskany podział zbioru obiektów należy poddać weryfikacji formalnej w celu określenia, na ile wyniki klasyfikacji odwzorowują rzeczywistą strukturę zjawiska. Wśród metod oceny jakości klasyfikacji w ADS można wymienić indeks sylwetkowy Rousseeuwa, analizę replikacji z indeksem Randa oraz metodę Bertranda i Bel-Mufti.

W wyniku klasyfikacji uzyskuje się informacje o liczbie i liczebności klas oraz przynależności obiektów do klas. W wielu badaniach istotne jest również rozpozna-

nie cech charakterystycznych i czynników różnicujących klasy. Klasy interpretuje się na podstawie zmiennych biorących udział w grupowaniu obiektów. W tym celu w ADS stosuje się technikę CLINT. W profilowaniu uczestniczą natomiast zmienne, które nie brały udziału w klasyfikacji, i zastosowanie mają metody statystycznej analizy wielowymiarowej. W ADS zaadaptowano metody analizy dyskryminacyjnej i drzew klasyfikacyjnych.

4. Oprogramowanie w klasyfikacji danych symbolicznych

Na rynku dostępnych jest wiele programów statystycznych wspomagających prowadzenie klasyfikacji, ale ich możliwości w zakresie ADS są zróżnicowane. Podstawowe cechy programów wraz ze wskazaniem możliwości tworzenia i wczytania tablicy danych symbolicznych oraz sporządzenia na jej podstawie macierzy odległości zaprezentowano w tab. 1.

Programem dedykowanym ADS, ale mało znanym w Polsce, jest SODAS. Instalacja programu wymaga klucza licencyjnego, który można bezpłatnie uzyskać od autorów. Daje on możliwość utworzenia lub wczytania tablicy danych symbolicznych oraz wyznaczenia na jej podstawie macierzy odległości. Program zawiera kilka podstawowych metod taksonomii symbolicznej, pozwala także dokonać interpretacji i profilowania klas.

R to środowisko do obliczeń statystycznych i jednocześnie język programowania działający w tym środowisku. Jest on programem bezpłatnym (również do zastosowań komercyjnych). Charakteryzuje się otwartym kodem źródłowym, co daje możliwość modyfikacji procedur i tworzenia własnych programów. Jego użytkowanie wymaga posiadania przynajmniej minimalnych umiejętności programistycznych. Umożliwia przeprowadzenie kompletnej procedury klasyfikacji.

Do popularnych w Polsce programów statystycznych należą SPSS i STATISTICA. Trudnością w stosowaniu tych programów w ADS, podobnie jak programu STATA, jest brak możliwości utworzenia tablicy danych symbolicznych oraz wyznaczenia na jej podstawie macierzy odległości. Z tego względu klasyfikacja danych symbolicznych jest możliwa jedynie po wczytaniu gotowej macierzy odległości. Na jej podstawie można przeprowadzić analizę z wykorzystaniem aglomeracyjnych metod taksonomii numerycznej.

Do zaawansowanych pakietów statystycznych należy również SAS. Podobnie jak program R daje on możliwość rozbudowy procedur, ale jest pakietem odpłatnym. Do tej pory nie oprogramowano żadnych funkcji pozwalających utworzyć, wczytać bądź wygenerować zbiór danych symbolicznych. Podobnie jak w przypadku programów SPSS, STATISTICA i STATA z dostępnych metod klasyfikacji (procedura CLUSTER) w ADS można skorzystać wtedy, gdy dysponuje się macierzą odległości obiektów symbolicznych.

Porównanie programów statystycznych pod względem przydatności w poszczególnych etapach procedury klasyfikacji obiektów symbolicznych zawiera tab. 2. Znak „+” oznacza, że w programie dostępne są funkcje przydatne w ADS.

Tabela 1. Podstawowe własności pakietów statystycznych

Lp.	Wyszczególnienie	Program					
		SODAS	R	SPSS	STATISTICA	STATA	SAS
1	Wydawca	FUNDP	R Foundation	IBM Corporation	StatSoft	Statacorp	SAS Institute
2	Dostęp (licencja)	Bezpłatny (ASSO)	Bezpłatny (GNU/ GPL)	Płatny	Płatny	Płatny	Płatny
3	Systemy operacyjne	Windows	Windows, Linux, MacOS	Windows, Linux, MacOS	Windows	Windows, Linux, MacOS	Windows, Linux, MacOS
4	Interfejs (język programowania)	Graficzny	Tekstowy (R)	Graficzny i tekstowy (Python, Sax Basic)	Graficzny	Tekstowy (Stata)	Graficzny i tekstowy (4GL, SQL)
5	Możliwość rozbudowy procedur	Nie	Tak	Tak	Nie	Tak	Tak
6	Dokumentacja/ podręcznik(i) w języku polskim	Nie/Nie	Nie/Tak	Tak/Tak	Tak/Tak	Nie/Tak	Tak/Tak
7	Obsługiwane formaty plików danych	xml, sds, mdb, xls	rda, csv, xml, sta, sas*	sav, xls, dbf, txt	sta, xls, dbf, csv, txt	dta, xls, txt	sas, xls, mdb, sav, txt
8	Możliwość utworzenia/ wczytania tablicy danych symbolicznych	Tak/Tak (format xml, mdb, xls)	Nie/Tak (format xml)	Nie/Nie	Nie/Nie	Nie/Nie	Nie/Nie
9	Możliwość utworzenia/ wczytania macierzy odległości obiektów symbolicznych	Tak/Nie	Tak/Tak	Nie/Tak	Nie/Tak	Nie/Tak	Nie/Tak

Źródło: opracowanie własne.

Tabela 2. Dostępność metod ADS w programach statystycznych

Lp.	Metody stosowane w procesie klasyfikacji	Program					
		SODAS	R	SPSS	STATISTICA	STATA	SAS
1	Metody wyboru zmiennych	–	+	–	–	–	–
2	Miary odległości	+	+	–	–	–	–
3	Metody klasyfikacji	+	+	+	+	+	+
4	Indeksy wyboru liczby klas	–	+	–	–	+	+
5	Metody oceny klasyfikacji	–	+	–	–	–	–
6	Metody interpretacji klas	+	+	–	–	–	–
7	Metody profilowania klas	+	+	–	–	–	–

Źródło: opracowanie własne.

5. Porównanie programów R i SODAS

Ze względu na relatywnie duże możliwości w zakresie ADS bliżej scharakteryzowane zostaną programy R i SODAS. SODAS składa się z panelu głównego oraz dwóch modułów, z których można skorzystać w poszczególnych etapach procedury klasyfikacyjnej (tab. 3).

Tabela 3. Moduły i opcje programu SODAS w klasyfikacji danych symbolicznych

Lp.	Wyszczególnienie	Moduł	Opcja	Funkcja
1	Zbiór danych symbolicznych	SOEDIT	–	–
2	Wybór obiektów i zmiennych	–	–	–
3	Pomiar odległości obiektów	Panel główny	Dissimilarity and Matching	DISS
4	Grupowanie obiektów	Panel główny	Clustering	DClust, SClust, HI_PYR
5	Wybór liczby klas	–	–	–
6	Ocena wyników klasyfikacji	–	–	–
7	Interpretacja klas	Panel główny	Clustering	CLINT
		VSTAR	–	–
8	Profilowanie klas	Panel główny	Discrimination & Regression	SCLASS

Źródło: opracowanie na podstawie [Noirhomme-Fraiture 2004a; 2004b].

Do utworzenia i edytowania tablicy danych symbolicznych służy moduł SOEDIT. Poza tym zbiór danych symbolicznych w plikach o formatach `xm1`, `mdb` lub `xls` można zaimportować przez interfejs ODBC.

Klasyfikację przeprowadza się w panelu głównym. Za pomocą opcji *Dissimilarity and Matching* oraz funkcji *DISS* można dokonać pomiaru odległości obiektów symbolicznych. Opcja *Clustering* umożliwia przeprowadzenie klasyfikacji obiektów symbolicznych z wykorzystaniem metod optymalizacyjnych (*SCLUST* i *DCLUST*) lub aglomeracyjnej metody Brito (*HI_PYR*). W opisie klas obiektów symbolicznych zastosowanie ma funkcja *CLINT* (moduł *Clustering*) oraz graficzny moduł *VSTAR*, w profilowaniu zaś funkcja *SCLASS* dostępna w opcji *Dissimilarity and Matching*. Program nie udostępnia metod selekcji zmiennych, indeksów wyboru liczby klas ani metod oceny wyników klasyfikacji.

Przeprowadzenie klasyfikacji w programie R wymaga załadowania odpowiednich pakietów oraz wczytania tablicy danych symbolicznych¹ w formacie *xml* za pomocą funkcji *parse.SO* bądź wygenerowania zbioru danych symbolicznych z wykorzystaniem funkcji *data_symbolic* lub *generate.SO*. Funkcje i pakiety mające zastosowanie w klasyfikacji danych symbolicznych zaprezentowano w tab. 4.

Tabela 4. Pakiety i funkcje programu R w klasyfikacji danych symbolicznych

Lp.	Wyszczególnienie	Pakiet	Funkcja
1	Zbiór danych symbolicznych	<i>clusterSim</i>	<i>data_symbolic</i>
		<i>symbolicDA</i>	<i>generate.SO</i> , <i>parse.SO</i>
2	Wybór zmiennych	<i>clusterSim</i>	<i>HINoV.Symbolic</i>
		<i>symbolicDA</i>	<i>HINoV.SDA</i> , <i>IchinoFS.SDA</i>
3	Pomiar odległości obiektów	<i>symbolicDA</i>	<i>dist.SDA</i>
		<i>clusterSim</i>	<i>dist.Symbolic</i>
4	Grupowanie obiektów	<i>cluster</i>	<i>diana</i> , <i>pam</i> , <i>agnes</i>
		<i>stats</i>	<i>hclust</i>
		<i>symbolicDA</i>	<i>DClust</i> , <i>SClust</i>
5	Wybór liczby klas	<i>clusterSim</i>	<i>index.G2</i> , <i>index.G3</i>
		<i>symbolicDA</i>	<i>index.G1d</i>
6	Ocena wyników klasyfikacji	<i>clusterSim</i>	<i>index.S</i>
		<i>symbolicDA</i>	<i>replication.SDA</i>
7	Interpretacja klas	<i>symbolicDA</i>	<i>cluster.Description.SDA</i>
8	Profilowanie klas	<i>symbolicDA</i>	<i>kernel.SDA</i> , <i>decissionTree.SDA</i>

Źródło: opracowanie na podstawie [Wilk 2011].

¹ Tablicę danych symbolicznych można utworzyć w programach *SODAS* i *SDAEditor*. Program *SDAEditor* wraz z instrukcją użytkownika jest dostępny na stronie www.ae.jgora.pl/keii (zakładka „Do pobrania”).

Do selekcji zmiennych metodą HINoV służą funkcje `HINoV.Symbolic` i `HINoV.SDA`, a w przypadku metody grafowej Ichino funkcja `IchinoFS.SDA`. Pomiaru odległości obiektów opisanych zmiennymi symbolicznymi można dokonać z wykorzystaniem funkcji `dist.Symbolic` i `dist.SDA`.

W pakiecie `symbolicDA` oprogramowano optymalizacyjne metody taksonomii symbolicznej (`SCLust` i `DClust`). Wiele bibliotek zawiera metody taksonomii numerycznej bazujące na macierzy odległości, które mają zastosowanie w ADS, np. `stats` (funkcja `hclust`, zawierająca metody hierarchiczne, m.in. metodę Warda – `ward`), `cluster` (metody hierarchiczne i optymalizacyjna metoda k -medoidów – `pam`).

Tabela 5. Korzyści i ograniczenia programów R i SODAS

Pakiet statystyczny	SODAS	R
Korzyści	<ul style="list-style-type: none"> – program bezpłatny, – podręcznik do ADS w programie SODAS, – oprogramowane metody taksonomii symbolicznej – możliwość wizualizacji wyników (porównania klas), – możliwość tworzenia tablicy danych symbolicznych lub importu przez interfejs ODBC, – dołączone zbiory danych symbolicznych 	<ul style="list-style-type: none"> – program bezpłatny, – podręcznik do ADS w programie R w języku polskim, – wiele pakietów związanych z klasyfikacją, – możliwość zrealizowania kompletnej procedury klasyfikacji danych symbolicznych, – możliwość rozbudowy procedur i dopasowania do potrzeb użytkownika, – możliwość generowania zbioru obiektów symbolicznych o zadanej strukturze klas, – dołączone zbiory danych symbolicznych
Ograniczenia	<ul style="list-style-type: none"> – brak możliwości zrealizowania kompletnej procedury klasyfikacji danych symbolicznych, – brak dokumentacji i podręcznika w języku polskim, – błędy w oprogramowaniu przejawiające się zawieszaniem lub zamykaniem programu 	<ul style="list-style-type: none"> – brak dokumentacji w języku polskim, – potrzeba posiadania przynajmniej minimalnych umiejętności programistycznych

Źródło: opracowanie własne.

Wśród dostępnych w programie indeksów wyboru liczby klas w ADS można zastosować indeks Bakera i Huberta (`index.G2`), Huberta i Levine (`index.G3`) i indeks Calińskiego i Harabasha (`index.G1d`). Na etapie oceny wyników klasyfikacji można posłużyć się analizą replikacji (`replication.SDA`) lub indeksem sylwetkowym (`index.S`). Oprogramowano również technikę CLINT (`cluster.Description.SDA`) służącą interpretacji klas, a także algorytm TREE z rodziny drzew klasyfikacyjnych (`decisionTree.SDA`) i jądrową analizę dyskryminacyjną – stosowane w profilowaniu klas obiektów symbolicznych (`kernel.SDA`).

Mocne i słabe strony obu pakietów w zakresie dostępności metod i technik służących klasyfikacji danych symbolicznych i własności użytkowych zestawiono w tab. 5.

6. Podsumowanie

Wśród najpopularniejszych w Polsce programów statystycznych umożliwiających przeprowadzenie klasyfikacji można wymienić pakiety SAS, SPSS, STATISTICA i R. W klasyfikacji danych symbolicznych najszersze zastosowanie znajdują obecnie programy R oraz SODAS. SODAS nie pozwala zrealizować kompletnej procedury klasyfikacyjnej, ale jeszcze kilka lat temu był jedynym programem dedykowanym analizie danych symbolicznych. Większe możliwości analityczne daje program R, w którym opracowywane są kolejne pakiety i funkcje poświęcone ADS. W przygotowaniu jest m.in. dedykowany ADS pakiet `clamix`. Ponadto w 2011 r. została wydana publikacja w polskiej wersji językowej na temat analizy danych symbolicznych w programie R.

Pakiety SPSS i STATISTICA dają bardzo ograniczone możliwości w klasyfikacji danych symbolicznych. Wymuszają bowiem korzystanie z metod bazujących na macierzy odległości, którą wcześniej należy wyznaczyć w programie R lub SODAS. Tworzony jest również komercyjny program SYROKKO (we francuskiej wersji językowej), w którym dostępny będzie moduł `ClustSYR`, służący klasyfikacji danych symbolicznych.

Literatura

- Anderberg M.R., *Cluster Analysis for Applications*, Academic Press Inc., New York 1973.
- Bock H.H., Diday E. (red.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin, Heidelberg 2000.
- Diday E., Noirhomme-Fraiture M. (red.), *Symbolic Data Analysis and the Sodas Software*, John Wiley & Sons, Chichester 2008.
- Everitt B.S., Landau S., Leese M., *Cluster Analysis*, 4th Edition, Arnold, London 2001.
- Gatnar E., *Symboliczne metody klasyfikacji danych*, PWN, Warszawa 1998.
- Grabiński T., Wydymus S., Zeliaś A., *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, PWN, Warszawa 1989.
- Noirhomme-Fraiture M. (red.), *Help Guide for SODAS 2 Software*, Software Report, Analysis System of Symbolic Official Data, Project no. IST-2000-2561, 2004a.
- Noirhomme-Fraiture M. (red.), *User Manual for SODAS 2 Software*, Software Report, Analysis System of Symbolic Official Data, Project no. IST-2000-25161, 2004b.
- Punj G., Stewart D.W., *Cluster analysis in marketing research: review and suggestions for application*, „Journal of Marketing Research” 1983, Mai, vol. 20.
- Verde R., *Clustering Methods in Symbolic Data Analysis*, [w:] D. Banks, L. House, E.R. McMorris, P. Arabie, W. Gaul (red.), *Classification, Clustering and Data Mining Applications*, Springer-Verlag, Heidelberg 2004.

- Walesiak M., *Problemy decyzyjne w procesie klasyfikacji zbioru obiektów*, [w:] J. Dziechciarz (red.), *Ekonometria 13, Zastosowania metod ilościowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1010, Wydawnictwo AE, Wrocław 2004.
- Wilk J., *Cluster Analysis Methods in Symbolic Data Analysis*, [w:] J. Pocięcha (red.), *Data Analysis Methods in Economic Investigations*, Studia i Prace UE w Krakowie nr 11, Kraków 2010.
- Wilk J., *Analiza skupień na podstawie danych symbolicznych*, [w:] E. Gatnar, M. Walesiak (red.), *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, PWN, Warszawa 2011.

COMPARATIVE STUDY OF SYMBOLIC DATA CLASSIFICATION SOFTWARE

Summary: The aim of this paper is to make an assessment of statistical software usefulness in symbolic data classification. In the paper the basic concepts of symbolic data analysis are explained and symbolic data classification procedure is discussed. Statistical packages for symbolic data classification are characterized and their usefulness in particular stages of classification procedure is identified. R and SODAS packages are described more precisely due to relatively wide application capabilities.

Keywords: classification, symbolic data, software.