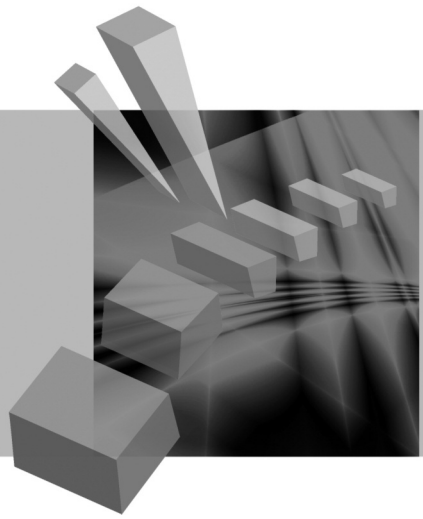
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

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 Pociecha,
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

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

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

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

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 Harabasza), 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	Dissimarity 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 `xml`, `mdb` lub `xls` można zaimportować przez interfejs ODBC.

Klasyfikację przeprowadza się w panelu głównym. Za pomocą opcji `Dissimarity 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 `Dissimarity 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 symbolicznych[1] 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	`clusterSim`	`data_symbolic`
		`symbolicDA`	`generate.SO`, `parse.SO`
2	Wybór zmiennych	`clusterSim`	`HINoV.Symbolic`
		`symbolicDA`	`HINoV.SDA`, `IchinoFS.SDA`
3	Pomiar odległości obiektów	`symbolicDA`	`dist.SDA`
		`clusterSim`	`dist.Symbolic`
4	Grupowanie obiektów	`cluster`	`diana`, `pam`, `agnes`
		`stats`	`hclust`
		`symbolicDA`	`DClust`, `SClust`
5	Wybór liczby klas	`clusterSim`	`index.G2`, `index.G3`
		`symbolicDA`	`index.G1d`
6	Ocena wyników klasyfikacji	`clusterSim`	`index.S`
		`symbolicDA`	`replication.SDA`
7	Interpretacja klas	`symbolicDA`	`cluster.Description.SDA`
8	Profilowanie klas	`symbolicDA`	`kernel.SDA`, `decissionTree.SDA`

Źródło: opracowanie na podstawie [Wilk 2011].

[1] Tablicę danych symbolicznych można utworzyć w programach SODAS i SDAEditor. Program SDAEditor wraz z instrukcją użytkowania 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	– 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	– 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	– brak możliwości zrealizowania kompletnej procedury klasyfikacji danych symbolicznych, – brak dokumentacji i podręcznika w języku polskim, – błędy w oprogramowaniu przejawiające się zawieszaniem lub zamykaniem programu	– 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 Harabasza (`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 (`decissionTree.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 SODS 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. Pociecha (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.