PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 86 2009
Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania

Małgorzata Gliwa
Akademia Ekonomiczna w Katowicach

WIZUALIZACJA OBIEKTÓW SYMBOLICZNYCH

1. Wstęp

W dobie dynamicznie rozwijających się technik komputerowych powstają coraz większe bazy danych. Często zdarza się, że dane zgromadzone w tych bazach mają postać danych symbolicznych. Ważna przy tym staje się umiejętność właściwego korzystania z tych baz danych, odkrywania wiedzy w nich zawartej, a także wizualizacji danych. W przypadku danych symbolicznych, które uważane są za nową jednostkę statystyczną, można wskazać kilka metod wizualizacji. Przykładem są mapy Kohonena, *Bi-dimensional mapping*, tabela obiektów symbolicznych, dendrogramy czy wykresy otrzymywane techniką *Zoom Star*. Jednak istotnym ograniczeniem w przypadku wielu metod wizualizacji obiektów symbolicznych jest wymóg dotyczący stosowania określonego typu zmiennych, co w przypadku rzeczywistych baz danych jest trudne do osiągnięcia. Uniwersalną w tym względzie techniką jest *Zoom Star*, która to technika daje możliwość graficznej prezentacji obiektów symbolicznych opisanych za pomocą zmiennej dowolnego typu.

Celem artykułu jest przedstawienie możliwości wizualizacji obiektów symbolicznych, w tym techniki *Zoom Star*. W części empirycznej przeprowadzona zostanie klasyfikacja obiektów symbolicznych pochodzących z rzeczywistego zbioru danych. Otrzymane wyniki przedstawione zostaną graficznie w formie zarówno dendrogramu, jak i wykresów dwuwymiarowych oraz trójwymiarowych. Na podstawie otrzymanych wykresów dokonana zostanie analiza porównawcza otrzymanych klas obiektów.

2. Zmienna symboliczna i obiekt symboliczny

Przedmiotem analizy danych symbolicznych (*Symbolic Data Analysis*) są obiekty, które są koniunkcją wartości poszczególnych jego cech. Cechy te są zmiennymi symbolicznymi, które mogą mieć postać [Bock, Diday 2000, s. 49]:

a) danych liczbowych,

b) przedziałów liczbowych rozłącznych lub posiadających część wspólną,

c) listy kategorii lub listy wartości,

d) listy kategorii lub listy wartości z wagami.

Ponadto według Didaya obiekty symboliczne mogą być również reprezentowane przez zmienne strukturalne [Diday 2002, s. 7].

Obiekty opisane przez zmienne symboliczne nazywane są obiektami symbolicznymi (*symbolic object*). Można określić dwa typy obiektów symbolicznych, a mianowicie obiekt symboliczny rzędu pierwszego (*first order symbolic object*) i obiekt symboliczny rzędu drugiego (*second order symbolic object*) [Bock, Diday 2000, s. 41]. Obiekt rzędu pierwszego charakteryzuje elementarną jednostkę (podobnie jak w klasycznej analizie danych), natomiast obiekt rzędu drugiego powstaje przez grupowanie obiektów rzędu pierwszego ze względu na określoną zmienną (zob. tab. 1 oraz tab. 2).

Tabela 1. Obiekty symboliczne pierwszego rzędu

Obiekty symboliczne	Zmienne			
	wykształcenie	płeć	dochód (zł)	stanowisko
Pracownik 1	wyższe	K	1200	pracownik_biu
Pracownik 2	średnie	M	2500	górnik
Pracownik 3	średnie	K	1100	sprzedawca
:	:	:	:	:

Źródło: opracowanie własne.

Tabela 2. Obiekty symboliczne drugiego rzędu pogrupowane według zmiennej wykształcenie

Obiekt symboliczny	Zmienne			
	wykształcenie	płeć	dochód (zł)	stanowisko
o_1	wyższe	{*K*(2/3, *M*(1/3)}	[1200, 2400]	{pracownik_biu (1/3), nauczyciel (1/3), lekarz (1/3)}
o_2	średnie	{*K*(1/2, *M*(1/1)}	[1100, 2500]	{górnik(1/2), sprzedawca(1/2)}
:	:	:	:	:

Źródło: opracowanie własne.

Diday wskazuje trzy źródła danych, na podstawie których tworzone są obiekty symboliczne [Diday 2002, s. 8], takie jak:

1) duże zbiory danych z danymi indywidualnymi,

2) tabele bazy danych zawierające informacje nie o danych indywidualnych, lecz o grupach danych,

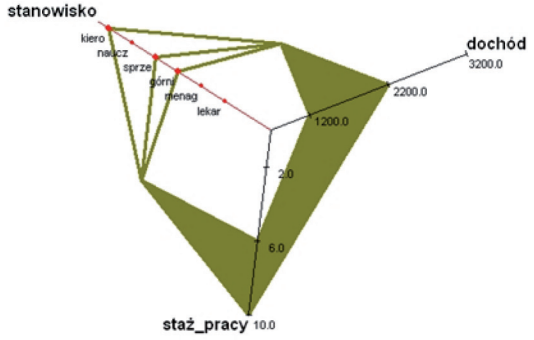
3) wiedza ekspertów.

3. Charakterystyka techniki *Zoom Star*

Technika *Zoom Star* pozwala na graficzną prezentację obiektów symbolicznych za pomocą wykresów dwuwymiarowych oraz trójwymiarowych. Idea *Zoom Star* odwołuje się do wykresów zwanych radarowymi. Każdej zmiennej odpowiada oś wartości, która promieniście odchodzi od punktu centralnego. Różnica pomiędzy wykresami radarowymi a wykresami otrzymywanymi w wyniku zastosowania techniki *Zoom Star* polega na tym, że w tym przypadku osiom wartości mogą być przyporządkowane nie tylko wartości liczbowe, ale również przedziały liczbowe, listy kategorii (wartości), listy kategorii (wartości) z wagami czy zmienne strukturalne [Noirhomme-Fraiture, Rouard 1998, s. 2; 2000, s. 126]. Ponadto dla zmiennych z wagami istnieje również możliwość uzyskania wykresu rozkładu wag [Noirhomme-Fraiture 2002, s. 4].

Do zalet tej metody należy możliwość porównywania charakterystyk kilku obiektów symbolicznych [Noirhomme-Fraiture 2002, s. 4-5], a także przedstawienia na jednym wykresie nie tylko wszystkich zmiennych charakteryzujących dany obiekt symboliczny, ale także wybranych zmiennych. Zamieszczone poniżej rys. 1-5 sporządzone zostały na podstawie danych umownych.

Rys. 1. Dwuwymiarowy wykres przedstawiający obiekt opisany przez wszystkie zmienne



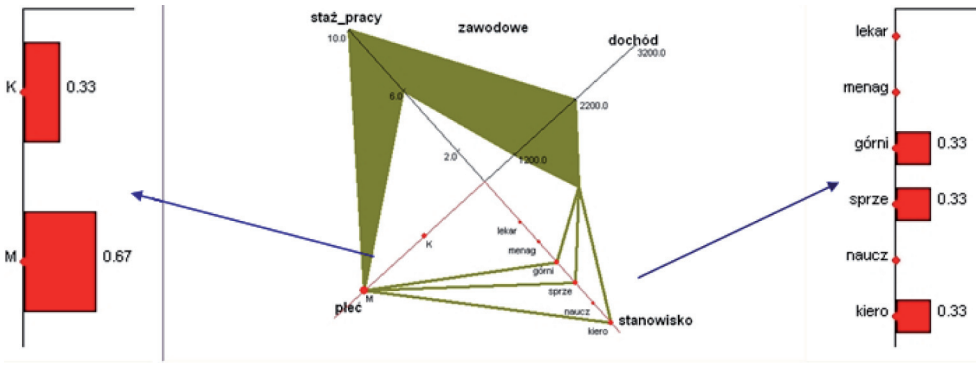
Rys. 2. Dwuwymiarowy wykres przedstawiający obiekt opisany przez wybrane zmienne

Źródło: opracowanie własne w programie SODAS.

Zoom Star to narzędzie wykorzystywane do graficznej prezentacji obiektów symbolicznych. Na podstawie powstałych wykresów odkrywać można relacje, związki zachodzące pomiędzy zmiennymi czy badanymi obiektami. Za pomocą techniki *Zoom Star* na wykresach można przedstawiać obiekty symboliczne zarówno z danych wejściowych, jak i te będące wynikami klasyfikacji.

W przypadku wykresów dwuwymiarowych na każdej osi zaznaczone są wartości poszczególnych zmiennych. Jeśli zmienna jest w postaci przedziału liczbowego,

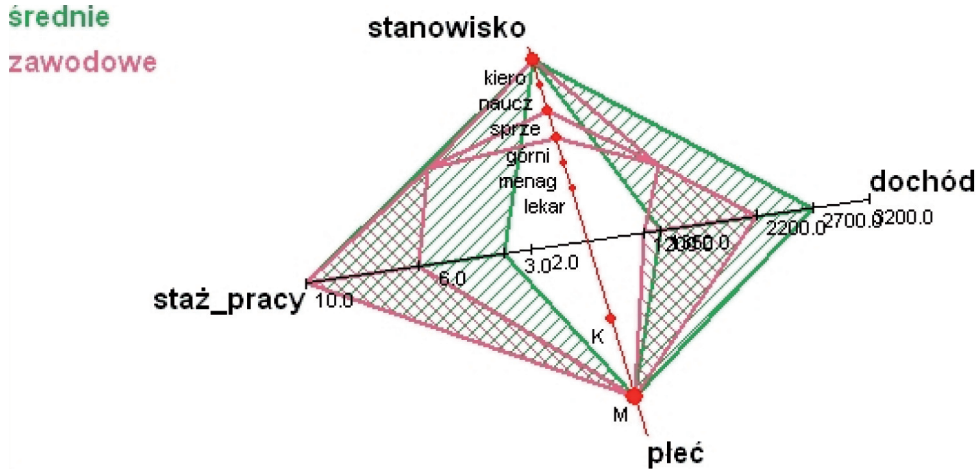
to wówczas na osi znajduje się dolna i górna granica przedziału, w przypadku listy kategorii lub wartości na osi pojawiają się punkty różnej wielkości, w zależności od wagi kategorii (wartości). Dodatkowo można otrzymać rozkład wag. Charakterystykę obiektu wyznaczają łamane łączące poszczególne wartości z osi zmiennych.



Rys. 3. Wykres dwuwymiarowy wraz z rozkładem wag

Źródło: opracowanie własne w programie SODAS.

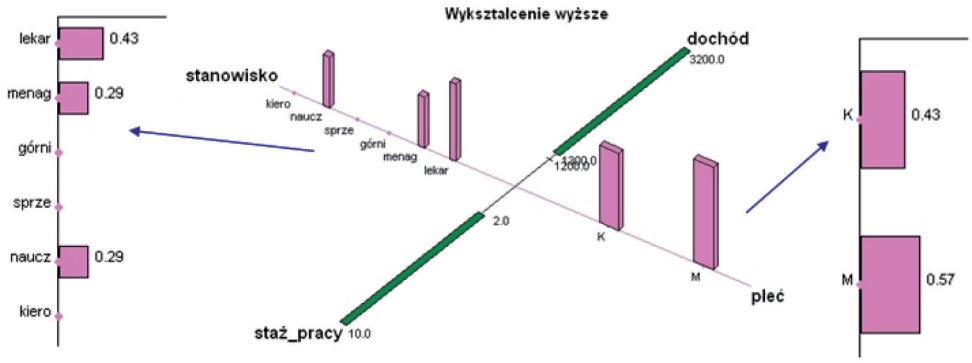
Wykresy dwuwymiarowe otrzymane techniką *Zoom Star* w łatwy sposób pozwalają na porównywanie obiektów symbolicznych. Uczynić to można przez wygenerowanie kilku wykresów obok siebie lub nałożenie tych wykresów na siebie [Noirhomme-Fraiture, Rouard 1998, s. 2] przez zastosowanie funkcji *Superimpose* programu SODAS.



Rys. 4. Porównanie dwóch obiektów przez zastosowanie funkcji *Superimpose*

Źródło: opracowanie własne w programie SODAS.

W przypadku wykresów trójwymiarowych dla zmiennej w postaci przedziału liczbowego na osi wartości pomiędzy dolną a górną granicą przedziału znajduje się zacieniowany obszar. Natomiast zmienne wyrażone w postaci listy kategorii lub listy wartości obrazowane są poprzez histogramy. Analogicznie, jak w przypadku wykresów dwuwymiarowych, dla zmiennych z wagami również można otrzymać rozkład wag.



Rys. 5. Wykres trójwymiarowy wraz z rozkładem wag

Źródło: opracowanie własne w programie SODAS.

Tak jak na wykresach dwuwymiarowych w ten sam sposób można porównywać charakterystyki kilku obiektów na wykresach trójwymiarowych [Noirhomme-Fraiture, Rouard 1998, s. 6].

Zarówno na wykresach dwuwymiarowych, jak i na wykresach trójwymiarowych można przedstawić zmienną hierarchiczną wraz z hierarchią zmiennych. Wówczas na wykresie, przy osi odpowiadającej tej zmiennej, znajduje się specjalny symbol [Noirhomme-Fraiture, Rouard 2000, s. 129-130].

4. Przykład empiryczny

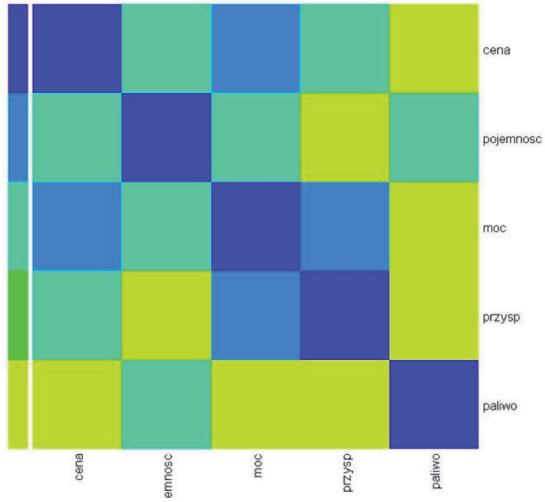
Celem poniższego przykładu jest przedstawienie wybranych metod wizualizacji danych i obiektów symbolicznych.

W przykładzie wykorzystano zbiór 12 obiektów symbolicznych scharakteryzowanych przez 5 zmiennych przedziałowych, takich jak: cena samochodu[1], pojemność silnika, moc, przyspieszenie, spalanie paliwa w cyklu miejskim.

Korelacje pomiędzy zmiennymi ilustruje rys. 6. Macierz korelacji pomiędzy zmiennymi symbolicznymi, stanowiącą podstawę do wykonania rysunku, wygene-

[1] Ceny samochodów pochodzą z opracowania na podstawie danych producentów z okresu 5.02.2008-22.04.2008.

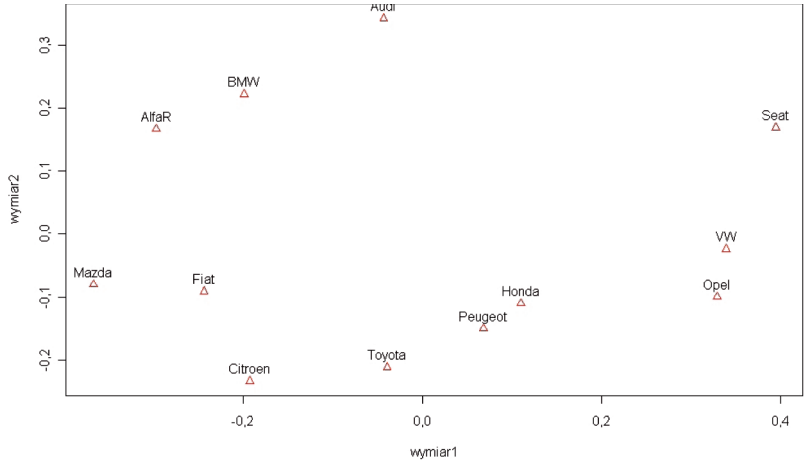
rowano za pomocą modułu DSTATS programu SODAS. Powstałą macierz przedstawiono graficznie z wykorzystaniem funkcji `heatmap(stats)` programu `R`. Im wyższa wartość współczynnika korelacji, tym większe natężenie koloru.



Rys. 6. Graficzna prezentacja macierzy współczynników korelacji

Źródło: opracowanie własne za pomocą programu R.

Na podstawie otrzymanej ilustracji wyszukiwać można zależności między zmiennymi, zwłaszcza w dużych zbiorach danych, gdzie analiza macierzy korelacji może stanowić problem.



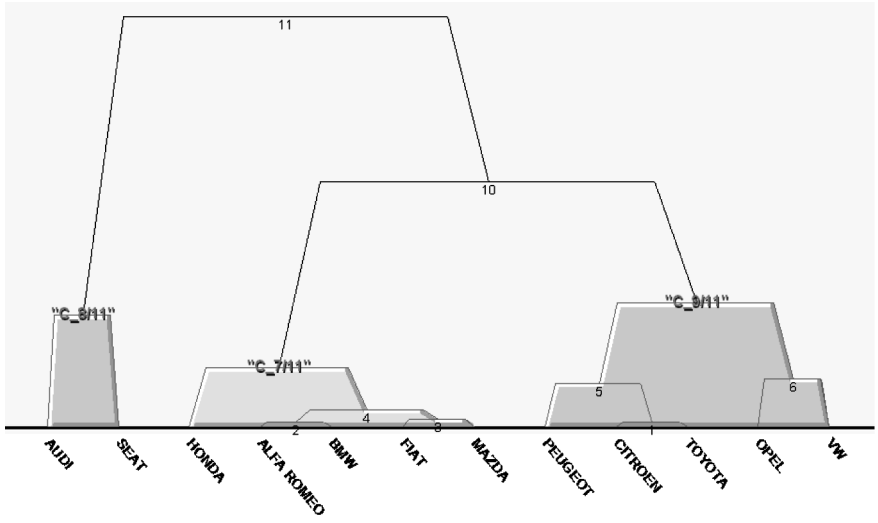
Rys. 7. Konfiguracja obiektów na płaszczyźnie

Źródło: opracowanie własne w programie R.

W kolejnym kroku przedstawiono konfigurację obiektów reprezentujących poszczególne marki samochodów. Rysunek 7 sporządzony został za pomocą analizy skalowania wielowymiarowego z wykorzystaniem funkcji `cmdscale(stats)` programu `R`.

Zadaniem skalowania wielowymiarowego jest przedstawienie w przestrzeni *r*-wymiarowej relacji zachodzących między badanymi obiektami [Gatnar, Walesiak 2004, s. 246]. Dla $r = 2$ odwzorowanie konfiguracji obiektów będzie na płaszczyźnie. Dane wejściowe, wykorzystane do sporządzenia rysunku, stanowiła macierz odległości pomiędzy obiektami symbolicznymi. Do wyznaczenia odległości zastosowano znormalizowaną odległość Ichino-Yaguchiego, która jest miarą stosowaną dla obiektów symbolicznych [Ichino, Yaguchi 1994, s. 698-704]. Obliczenia wykonane zostały w programie SODAS.

W kolejnej części przeprowadzono grupowanie obiektów symbolicznych za pomocą hierarchicznej metody aglomeracyjnej dla zbioru obiektów symbolicznych. W zastosowanej metodzie grupowanie obiektów odbywa się nie na podstawie macierzy odległości, lecz jest przeprowadzane bezpośrednio dla zbioru obiektów symbolicznych. Otrzymane wyniki klasyfikacji zostały graficznie przedstawione w postaci dendrogramu.



Rys. 8. Dendrogram dla zbioru obiektów symbolicznych

Źródło: opracowanie własne w programie SODAS.

W wyniku przeprowadzonej klasyfikacji otrzymano klasy obiektów, które są jednorodne ze względu na określone kryterium. Każda z otrzymanych klas reprezentowana jest przez jeden obiekt symboliczny, zwany zupełnym. W klasie 1, oznaczonej na rys. 8 jako C_8/11, znalazły się następujące obiekty symboliczne: {Audi, Seat}. Klasa ta reprezentowana jest przez następujący zupełny obiekt symboliczny:

```
[cena = [54990, 130435]] ^ [pojemnosc = [1595, 1984]] ^
[moc = 102, 240]] ^ [przysp/100 = [6.6, 17]] ^ [paliwo_miasto =
[5, 11.7]].
```

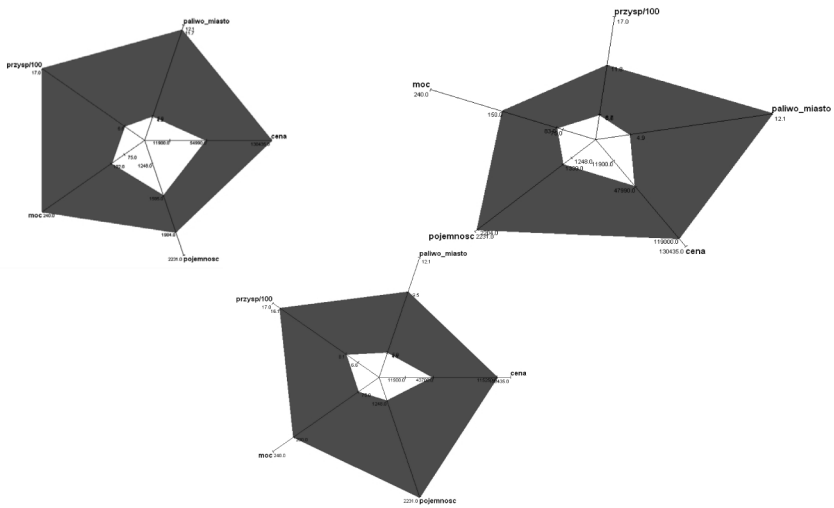
Klasa 2, oznaczona jako C_7/11, zawiera obiekty symboliczne: {Honda, Alfa Romeo, BMW, Fiat, Mazda} i jest reprezentowana przez następujący zupełny obiekt symboliczny:

```
[cena = [47990, 119000]] ^ [pojemnosc = [1339, 2204]] ^
[moc = [83, 150]] ^ [przysp/100 = [6.7, 11.8]] ^ [paliwo_miasto =
[4.9, 12.1]].
```

W klasie 3, oznaczonej jako C_9/11, znalazły się obiekty: {Peugeot, Citroen, Toyota, Opel, VW}. Klasę tę charakteryzuje zupełny obiekt symboliczny zapisany w poniższej formie:

```
[cena = [43700, 115250]] ^ [pojemnosc = [1248, 2231]] ^
[moc = [75, 200]] ^ [przysp/100 = [8.1, 16.1]] ^ [paliwo_miasto =
[5, 9.5]].
```

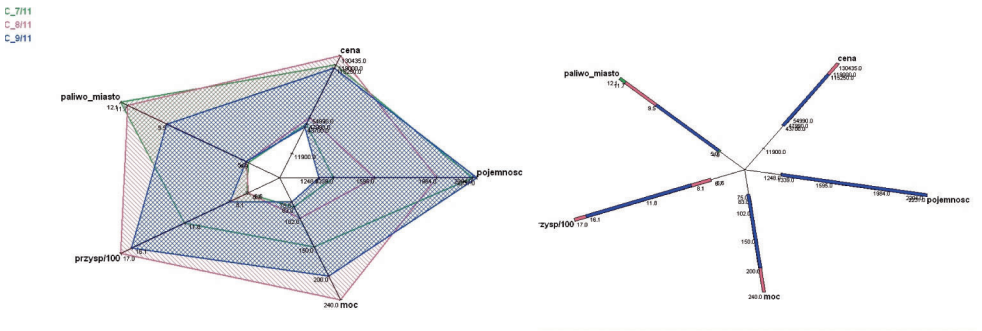
Oprócz charakterystyki obiektów w postaci koniunkcji wartości cech obiektu otrzymane klasy można graficznie przedstawić za pomocą wykresów radarowych.



Rys. 9. Dwuwymiarowe wykresy klas obiektów symbolicznych

Źródło: opracowanie własne w programie SODAS.

Aby dokonać analizy porównawczej otrzymanych klas obiektów, klasy te można przedstawić na jednym wspólnym wykresie zarówno dwuwymiarowym, jak i trójwymiarowym. W tym celu należy zastosować funkcję *Superimpose*.



Rys. 10. Porównanie klas obiektów symbolicznych na wykresach: dwuwymiarowym i trójwymiarowym

Źródło: opracowanie własne w programie SODAS.

Powyższe rysunki (rys. 9 i 10) wykonane zostały z wykorzystaniem techniki *Zoom Star*.

Na podstawie otrzymanych wykresów można stwierdzić, że w klasie 1 znajdują się samochody o najwyższej mocy silnika, jednak są to samochody najdroższe. Do klasy 2 należą samochody ze „średniej" półki cenowej, jednak o najwyższym spalaniu paliwa w cyklu miejskim, natomiast do 3 należą samochody najtańsze i najbardziej ekonomiczne pod względem spalania paliwa w cyklu miejskim.

5. Podsumowanie

W artykule zaprezentowano wybrane metody graficznej prezentacji obiektów symbolicznych. Otrzymane wykresy pozwalają nie tylko na wizualizację obiektów, ale również na odkrywanie zależności między zmiennymi symbolicznymi. Przedstawiono możliwości wykorzystania techniki *Zoom Star* do wizualizacji obiektów symbolicznych. Jest to uniwersalne narzędzie, które może być stosowane niezależnie od typu zmiennych i pozwala na tworzenie wykresów zarówno 2D, jak i 3D. W kolejnej części przedstawiono graficzną prezentację macierzy korelacji, która może być użyteczna do wyszukiwania zależności pomiędzy zbiorami wielu zmiennych w dużych zbiorach danych. Pokazano również, w jaki sposób przedstawić na płaszczyźnie konfigurację obiektów symbolicznych.

Prezentowane w artykule wykresy to uproszczony, lecz czytelny sposób wizualizacji dużych zbiorów danych.

Wszystkie obliczenia oraz ilustracje wykonano za pomocą programu SODAS ver. 2.0 oraz R ver. 2.9.0.

Literatura

Bock H.H., Diday E., *Analysis of Symbolic Data*, Springer-Verlag, Berlin 2000, s. 49.

Diday E., *An Introduction to Symbolic Data Analysis and the Sodas Software,* „The Electronic Journal of Symbolic Data Analysis" 2002, nr 0.

Gatnar E., Walesiak M., *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław 2004, s. 246.

Ichino M., Yaguchi H., *Generalized Minkowski Metrics for Mixed Feature Type Data Analysis*, „IEEE Transactions on System, Man and Cybernetics" 1994, t. 24 (4), s. 698-704.

Noirhomme-Fraiture M., Rouard M., *Representation of Sub-Populations and Correlation with Zoom Star*, Proceedings of NTTS'98, EUSTAT, Sorrento1998.

Noirhomme-Fraiture M., Rouard M., *Visualizing and Editing Symbolic Object*, [w:] *Analysis of Symbolic Data*, H.-H. Bock, E. Diday (red.), Springer Verlag, Berlin-Heidelberg 2000, s. 125-138.

Noirhomme-Fraiture M., *Visualization of Large Data Sets. The Zoom Star Solution*, „The Electronic Journal of Symbolic Data Analysis" 2002, t. 0, nr 0.

VISUALIZATION OF SYMBOLIC OBJECTS

Summary

In the article, Zoom Star technique is presented, which allows for graphical representation of symbolic objects. 2D and 3D graphs were used. Next, classification of symbolic objects taken from real database was carried out. The hierarchical agglomerative method for symbolic objects was used. Results were presented both in the form of dendrogram and 2D and 3D graphs. On the basis of these graphs a comparative analysis of classes was done.

Calculations were made in the programme SODAS and R.