Alicja Grześkowiak
Uniwersytet Ekonomiczny we Wrocławiu

WSPOMAGANIE ANALIZ PROWADZONYCH W OPARCIU O ZMIENNE MIERZONE NA RÓŻNYCH SKALACH ZA POMOCĄ TECHNIK WIZUALIZACYJNYCH[1]

Wprowadzenie

W badaniach społecznych często występują zbiory zmiennych mierzonych na różnych skalach, przy czym znaczącą rolę pełnią dane o charakterze niemetrycznym. W literaturze przedmiotu proponuje się różne metody analityczne uwzględniające istnienie danych o mieszanym charakterze. Obecnie coraz ważniejszą rolę w analizie danych pełnią techniki wizualizacyjne, które nie tylko uzupełniają prezentację wyników badań, ale także umożliwiają eksplorację zbiorów i wykrywanie istniejących prawidłowości. Towarzyszy temu rozwój oprogramowania statystycznego w zakresie grafiki, zarówno w pakietach komercyjnych, jak i niekomercyjnych. Szczególnie bogate zasoby procedur zawierają biblioteki programu R.

Niniejszy artykuł traktuje o wybranych, relatywnie rzadko stosowanych, technikach wizualizacji przydatnych w przypadku dysponowania zmiennymi, których pomiaru dokonano na różnych skalach pomiarowych. Głównym celem jest przegląd procedur dostępnych w pakietach statystycznych wspomagających prowadzenie analiz na podstawie zbiorów danych o mieszanym charakterze. Rozważania dotyczą trzech obszarów:

– przedstawiania prawidłowości występujących w zbiorze danych,
– ilustrowania podobieństwa obiektów,
– prezentowania wyników wybranych analiz wielowymiarowych.

[1] Praca naukowa sfinansowana ze środków Narodowego Centrum Nauki w ramach projektu badawczego 2012/05/B/HS4/02499.

Jako materiał ilustracyjny zostały wykorzystane dane zebrane w ramach badania *Diagnoza społeczna*[2] w 2013 roku, odnoszące się do robienia zakupów przez Internet oraz charakteryzujące respondentów (N = 13 825). Zestaw rozpatrywanych zmiennych był następujący:

- kupowanie produktów i usług przez Internet: tak, nie (skala nominalna),
- płeć: kobieta, mężczyzna (skala nominalna),
- wykształcenie: podstawowe, zawodowe/gimnazjum, średnie, wyższe (skala porządkowa),
- wiek w latach (skala ilorazowa),
- czas korzystania z Internetu w ostatnim tygodniu w godzinach (skala ilorazowa).

Do wykonania wizualizacji zastosowano programy R, SPSS oraz Parallel Sets, przy czym z programu R wykorzystano różne pakiety, które wymieniono bezpośrednio przy omawianiu poszczególnych metod.

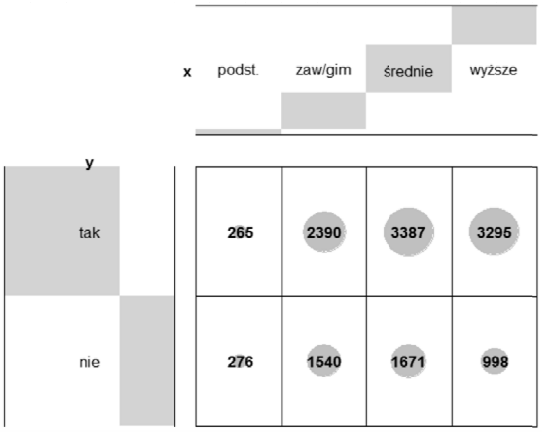
1. Graficzne przedstawianie prawidłowości w zbiorach danych

Relacje pomiędzy dwoma zmiennymi: jednej o charakterze nominalnym i drugiej o charakterze porządkowym można przedstawić, wizualizując odpowiadającą im tabelę kontyngencji, np. za pomocą wykresu bąbelkowego dostępnego w pakiecie `ggplot2`, na którym wielkości powierzchni kół odpowiadają liczebnościom pól tabeli[3], bądź stosując funkcję `balloonplot` z pakietu `gplots`, która dodatkowo uwzględnia cieniowanie odzwierciedlające liczebności brzegowe (rys. 1).

Na podstawie rys. 1 łatwo zauważyć, że w próbie dominowały osoby dokonujące zakupów przez Internet. Relatywnie najmniej respondentów legitymowało się wykształceniem podstawowym i wśród nich liczebność kupujących i niekupujących była bardzo zbliżona. Natomiast największą rozbieżność odnotowano wśród osób z wykształceniem wyższym, wśród których wyraźnie przeważają osoby deklarujące wykonywanie zakupów za pośrednictwem sieci.

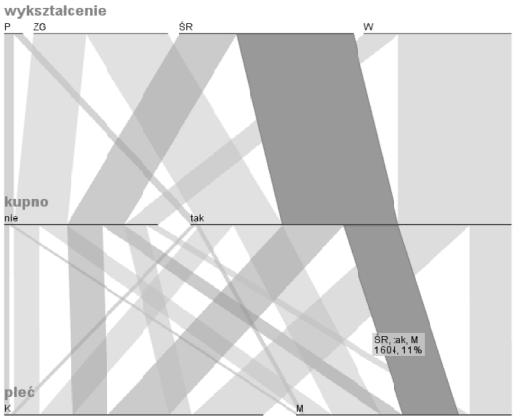
[2] Rada Monitoringu Społecznego (2013): Diagnoza społeczna: zintegrowana baza danych. www.diagnoza.com [28.03.2014].

[3] Zob. H. Wickham: ggplot2: elegant graphics for data analysis. Springer, New York 2009.



Rys. 1. Wizualizacja tabeli kontyngencji dla dwóch zmiennych niemetrycznych (kupno, wykształcenie) za pomocą wykresu bąbelkowego z uwzględnieniem liczebności brzegowych

Idea prezentacji przedstawiona na rys. 1 ma wiele zalet, ale ogranicza się do przypadku dwuwymiarowego. Chcąc zilustrować relacje pomiędzy większą liczbą zmiennych niemetrycznych, można posłużyć się wykresami zbiorów równoległych (*parallel sets*), na których połączenia pomiędzy kategoriami reprezentują liczbę kombinacji atrybutów[4]. Przykład wizualizacji trzech zmiennych: wykształcenia, faktu zakupu oraz płci przedstawia rys. 2.



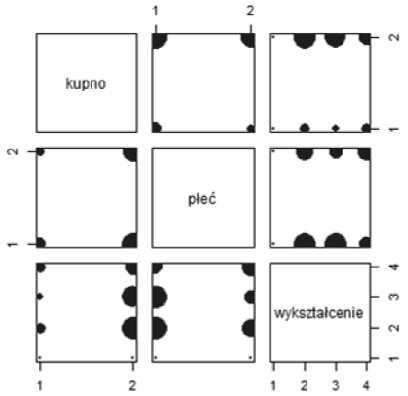
Rys. 2. Wizualizacja relacji pomiędzy trzema zmiennymi niemetrycznymi (wykształcenie, kupno, płeć) za pomocą wykresu zbiorów równoległych*

* Wizualizację uzyskano za pomocą programu *Parallel Sets.*

[4] R. Kosara, F. Bendix, H. Hauser: Parallel sets: Interactive exploration and visual analysis of categorical data. „Transactions on Visualization and Computer Graphics" 2006, Vol. 12, No. 4, s. 558-568.

Pierwszy podział odpowiada danym zawartym w dwuwymiarowej tabeli kontyngencji (analogicznie jak na rys. 1), drugi wnosi informacje o kolejnym wymiarze danych – płci, np. nietrudno zauważyć, że wśród respondentów o wykształceniu średnim niedokonujących zakupów przez Internet przeważają kobiety. Program *Parallel Sets* ułatwia również identyfikację poszczególnych kategorii. Przykładowo na rys. 2 wyróżniono najciemniejszym odcieniem mężczyzn o wykształceniu średnim wykorzystujących globalną sieć do zakupów. Takich jednostek jest 1604, co stanowi 11% badanych.

Kolejną propozycją ilustracji powiązań zmiennych mierzonych na słabych skalach jest wykres rozrzutu dla danych niemetrycznych, dostępny w R w pakiecie `clusterSim`[5]. Na tego rodzaju wykresie koła reprezentują pary kategorii, a promień częstość ich występowania (rys. 3).



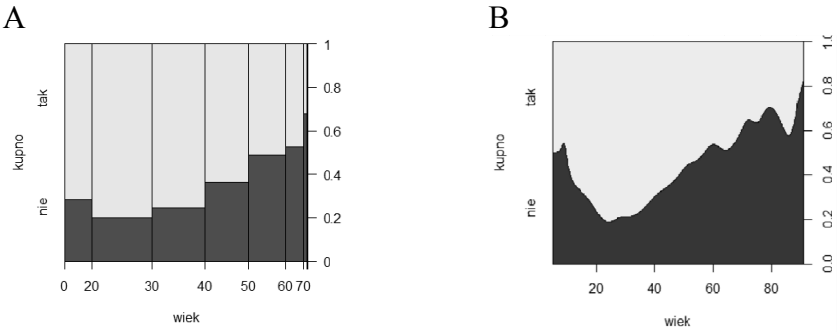
Rys. 3. Wykres rozrzutu dla trzech zmiennych niemetrycznych: kupno: 1 – nie, 2 – tak; płeć: 1 – kobieta, 2 – mężczyzna; wykształcenie: 1 – podstawowe i niższe, 2 – zawodowe/gimnazjum, 3 – średnie, 4 – wyższe i policealne

W pakiecie `clusterSim` dostępna jest także funkcja umożliwiająca tworzenie wykresu rozrzutu dla trzech zmiennych o charakterze niemetrycznym w przestrzeni trójwymiarowej, przy czym prezentacja przyjmuje formę kul, których promień odzwierciedla częstość występowania kategorii[6]. Tworzony wykres ma interaktywny charakter umożliwiający wgląd w strukturę danych.

[5] M. Walesiak, A. Dudek: clusterSim: Searching for optimal clustering procedure for a data set, R package ver. 0.43-4, 2014, http://cran.r-project.org/web/packages/clusterSim/index.html.

[6] Statystyczna analiza danych z wykorzystaniem programu R. Red. M. Walesiak, E. Gatnar. Wydawnictwo Naukowe PWN, Warszawa 2012, s. 102; Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R. Red. M. Walesiak, E. Gatnar. Wydawnictwo C.H. Beck, Warszawa 2011, s. 43-45.

Kolejną interesującą kwestią jest graficzne, łączne ujmowanie zmiennych metrycznych i niemetrycznych. W celu badania zależności jednej zmiennej niemetrycznej i jednej metrycznej można zastosować spinogram lub wykres warunkowej gęstości[7] przedstawione na rys. 4, a wykonane za pomocą funkcji dostępnych w R w pakiecie `vcd`[8].



Rys. 4. Ilustracja zależności zmiennej kupno od zmiennej wiek za pomocą spinogramu (A) i wykresu warunkowej gęstości (B)

W wizualizacji za pomocą spinogramu szerokość słupków odpowiada częstościom cechy metrycznej (wiek), a wysokość elementów cieniowanych obrazuje warunkową częstość cechy niemetrycznej (kupno). Wykres warunkowej gęstości oddaje tę samą zależność, ale w sposób wygładzony. Obie prezentacje wskazują na ogólną prawidłowość, że wraz z wiekiem maleje częstość robienia zakupów przez Internet, niesprawdzająca się jedynie w przypadku najmłodszej grupy wieku, być może ze względu na ograniczone środki finansowe.

Wykresy przedstawione na rys. 4 ukazują relacje tylko pomiędzy dwoma zmiennymi. Zależności pomiędzy większą ich liczbą można zilustrować za pomocą warunkowego wykresu rozrzutu (funkcja `coplot`) z pakietu `graphics` (adekwatnego dla dwóch zmiennych metrycznych i dwóch niemetrycznych) oraz uogólnionego wykresu rozrzutu, którego różne wersje oferują pakiety `GGally`[9] oraz `gpairs`[10]. Uogólniony wykres rozrzutu stanowi analogon tradycyjnego macierzowego wykresu rozrzutu stosowanego dla zmiennych metrycznych,

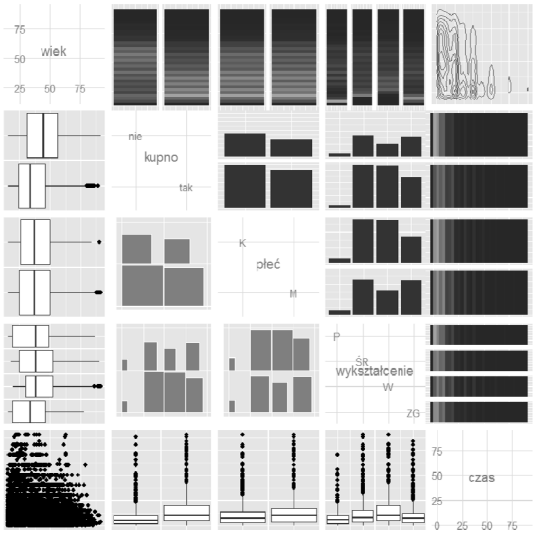
[7] Zob. B.S. Everitt, T. Hothorn: A handbook of statistical analyses using R. CRC Press, Boca Raton 2009, s. 37-38.

[8] D. Meyer, A. Zeileis, K. Hornik: vcd: Visualizing Categorical Data. R package ver. 1.3-1, 2013, http://cran.r-project.org/web/packages/vcd/index.html.

[9] B. Schloerke et al.: GGally: Extension to ggplot2, R package, ver. 0.4.6, 2014, http://cran.r-project.org/web/packages/GGally/index.html.

[10] W. Emerson, W.A. Green: gpairs: The Generalized Pairs Plot, R package ver 1.2, 2013, http://cran.r-project.org/web/packages/gpairs/index.html.

a rodzaj stosowanej prezentacji danych zależy od typu kombinacji: dla par zmienna metryczna – zmienna metryczna stosowane są wykresy rozrzutu z prezentacją wartości współczynnika korelacji oraz wykresy gęstości, dla par zmienna niemetryczna – zmienna niemetryczna wykresy mozaikowe, kolumnowe, fluktuacyjne, a dla par zmienna metryczna – zmienna niemetryczna wykresy pudełkowe i paskowe[11]. Wizualizacje można tworzyć symetrycznie lub korzystać z innych metod nad i pod główną przekątną. Na rys. 5 zaprezentowano przykładowy uogólniony wykres rozrzutu dla wszystkich rozpatrywanych zmiennych wykonany z użyciem pakietu `GGally`. W celu zaprezentowania różnorakich możliwości ilustracji zastosowano różne techniki wizualizacji pod (wykres rozrzutu, wykresy pudełkowe, wykresy fluktuacyjne) i nad główną przekątną (wykresy paskowe, wykresy kolumnowe, wykres gęstości).



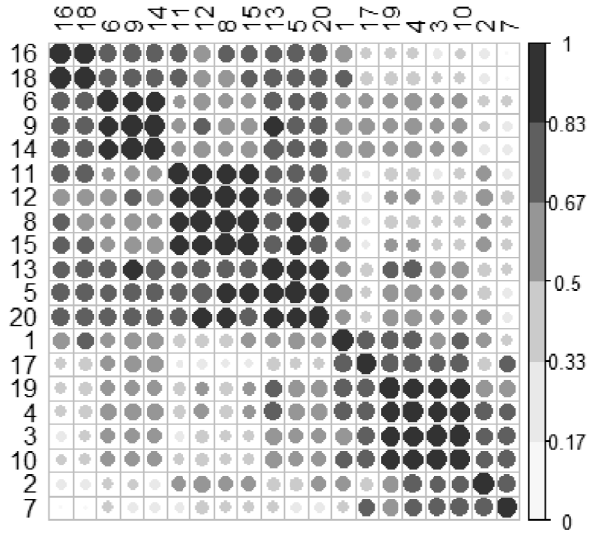
Rys. 5. Wizualizacja zależności pomiędzy pięcioma zmiennymi o różnym charakterze (wiek, kupno, płeć, wykształcenie, czas)

2. Obrazowanie podobieństwa obiektów

Jednym z celów analiz wielowymiarowych może być badanie stopnia podobieństwa obiektów opisanych za pomocą wielu cech. Pomiar podobieństwa nie stanowi wyzwania, gdy wszystkie zmienne są mierzone na tej samej skali,

[11] J.W. Emerson et al.: The generalized pairs plot. „Journal of Computational and Graphical Statistics" 2013, Vol. 22(1), s. 79-91.

lecz znacznie komplikuje się w sytuacji posiadania zbioru zawierającego zmienne mierzone na różnych skalach – w literaturze można odnaleźć kilka propozycji mierzenia odległości w tego rodzaju okolicznościach[12]. Dysponując macierzą odległości, można dokonać jej transformacji w macierz podobieństw. Obydwa typy macierzy można wizualizować za pomocą zróżnicowanej kolorystyki i symboli[13], a dodatkowo wykonać grupowanie obiektów podobnych. Ze zbioru danych dotyczących zakupów przez Internet wylosowano dwadzieścia obiektów, oceniono ich niepodobieństwo za pomocą miary Gowera, wyznaczono macierz podobieństwa, wykonano grupowanie jednostek, korzystając z hierarchicznej procedury aglomeracyjnej, a całościowy efekt zilustrowano graficznie, wykorzystując pakiet `corrplot`[14] progamu R (rys. 6).



Rys. 6. Ilustracja macierzy podobieństwa obiektów

Wizualizacja przedstawiona na rys. 6 ułatwia percepcję prawidłowości zachodzących w macierzy podobieństwa za pomocą natężenia koloru oraz wielkości symboli, a także pozwala wskazać grupy obiektów podobnych.

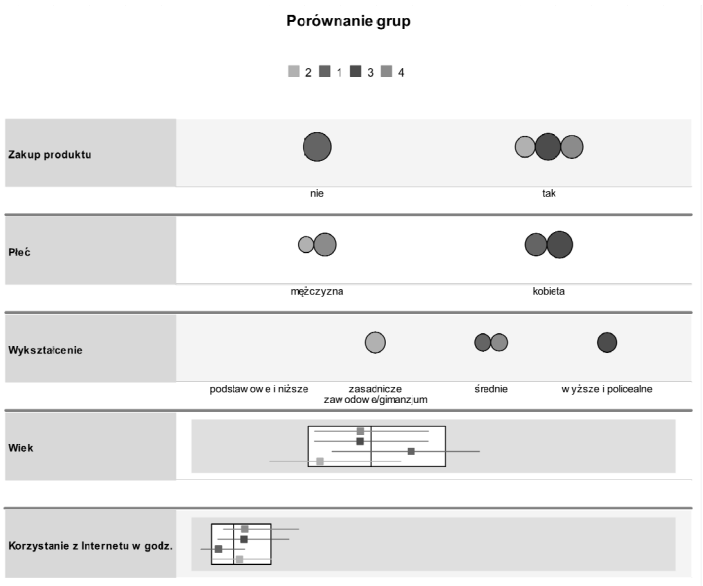
[12] M. Walesiak: Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R. UE we Wrocławiu, Wrocław 2011, s. 22-34.

[13] M. Friendly: Corrgrams: Exploratory displays for correlation matrices. „The American Statistician" 2002, Vol. 56 (4), s. 316-324.

[14] T. Wei: corrplot: Visualization of a correlation matrix, R package ver. 0.73, 2013, http://cran.r-project.org/web/packages/corrplot/index.html.

3. Wizualizacja wyników wybranych metod wielowymiarowych – grupowania dwustopniowego oraz analizy kanonicznej

Do metod wielowymiarowych umożliwiających jednoczesną analizę zmiennych różnie skalowanych należą grupowanie dwustopniowe oraz analiza kanoniczna dla zmiennych niemetrycznych. W grupowaniu dwustopniowym pozwalającym na określenie klas obiektów podobnych, oprogramowanym w SPSS, pierwszy etap polega na tworzeniu drzewa klasyfikacyjnego, natomiast w drugim tworzone są skupienia za pomocą hierarchicznej procedury aglomeracyjnej, a zestaw możliwych rozwiązań podlega ocenie za pomocą bayesowskiego kryterium informacyjnego[15]. Otrzymane grupy można porównać na podstawie reprezentacji graficznej ujmującej zmienne profilujące (rys. 7).

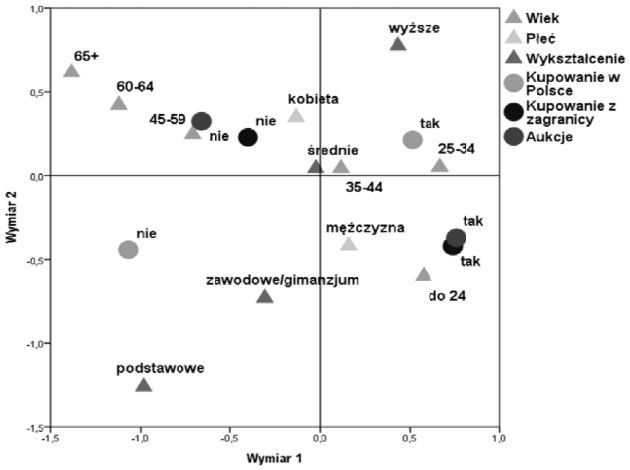


Rys. 7. Przedstawienie wyników grupowania dwustopniowego – podział na cztery grupy

Dla zmiennych niemetrycznych przedstawiana jest dominanta, natomiast metryczne są prezentowane za pomocą wykresów pudełkowych. Przykładowo, w grupie oznaczonej nr 1 (przewaga niedokonujących zakupów przez Internet) występuje dominacja kobiet, osób z wykształceniem średnim, o najwyższej medianie wieku i spędzających przeciętnie najmniej czasu korzystając z Internetu.

[15] M. Rószkiewicz: Analiza klienta. SPSS Polska, Kraków 2011, s. 82.

Techniką umożliwiającą badanie relacji pomiędzy kilkoma zbiorami zmiennych zmierzonych na różnych skalach jest analiza kanoniczna dla zmiennych niemetrycznych (OVERALS), a prezentacja graficzna rezultatów za pomocą środków ciężkości pozwala określić współwystępowanie kategorii[16]. Na rys. 8 przedstawiono wyniki analizy przeprowadzonej dla dwóch zbiorów zmiennych. Jeden z nich tworzyły zachowania konsumpcyjne (kupowanie przez Internet w Polsce, kupowanie przez Internet z zagranicy, uczestniczenie w aukcjach przez Internet), a drugi zestaw czynniki socjodemograficzne (płeć, wiek, wykształcenie).



Rys. 8. Wyniki analizy kanonicznej dla dwóch zbiorów zmiennych

Układ kategorii na rys. 8 wskazuje, że kupowanie przez Internet z zagranicy i na aukcjach to domena młodszych mężczyzn. Kobiety i osoby powyżej 45. roku życia są mniej skłonne do tego typu zachowań konsumpcyjnych. Natomiast zmienna wykształcenie w znacznej mierze warunkuje dokonywanie zakupów przez Internet w Polsce.

Podsumowanie

Badania społeczne często zasadzają się na zmiennych różnie skalowanych, przy czym znaczącą rolę pełnią dane o charakterze niemetrycznym. W eksploracji zbiorów danych oraz w przedstawianiu wyników badań coraz większe znaczenie zyskują techniki wizualizacji. O ile ukazywanie w sposób graficzny relacji dla zmiennych metrycznych jest powszechnie stosowane, np. za pomocą

[16] Zob. J.J. Meulman, W.J. Heiser: SPSS Categories 11.0. SPSS, Chicago 2001.

wykresów korelacyjnych, to ukazywanie powiązań dla zmiennych niemetrycznych lub różnie skalowanych wydaje się niedoceniane na gruncie polskich badań społecznych. Może to wynikać z faktu, że przedstawione w artykule rozwiązania są stosunkowo nowe i niezbyt rozpowszechnione. Jednocześnie warto podkreślić, że opisany zestaw narzędzi stanowi subiektywną selekcję zorientowaną na zaprezentowanie różnych metod w odniesieniu do trzech różnych celów analitycznych. W opinii autorki ułatwiają one i wzbogacają analizę danych i z tego powodu zasługują na rozpropagowanie i szerokie zastosowanie.

Literatura

Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R. Red. M. Walesiak, E. Gatnar. Wydawnictwo C.H. Beck, Warszawa 2011.

Emerson J.W., Green W.A., Schloerke B., Crowley J., Cook D., Hofmann H., Wickham H.: The generalized pairs plot. „Journal of Computational and Graphical Statistics" 2013, Vol. 22(1).

Emerson W., Green W.A.: gpairs: The Generalized Pairs Plot, R package ver 1.2, 2013, http://cran.r-project.org/web/packages/gpairs/index.html.

Everitt B.S., Hothorn T.: A handbook of statistical analyses using R. CRC Press, Boca Raton 2009.

Friendly M.: Corrgrams: Exploratory displays for correlation matrices. „The American Statistician" 2002, Vol. 56 (4).

Kosara R., Bendix F., Hauser H.: Parallel sets: Interactive exploration and visual analysis of categorical data. „Transactions on Visualization and Computer Graphics" 2006, Vol. 12, No. 4.

Meulman J.J., Heiser W.J.: SPSS Categories 11.0. SPSS, Chicago 2001.

Meyer D., Zeileis A., Hornik K.: vcd: Visualizing Categorical Data. R package ver. 1.3-1, 2013, http://cran.r-project.org/web/packages/vcd/index.html**.**

Rószkiewicz M.: Analiza klienta. SPSS Polska, Kraków 2011.

Schloerke B., Crowley J., Cook D., Hofmann H., Wickham H., Briatte F., Marbach M.: GGally: Extension to ggplot2, R package, ver. 0.4.6, 2014, http://cran.r-project.org/web/packages/GGally/index.html.

Statystyczna analiza danych z wykorzystaniem programu R. Red. M. Walesiak, E. Gatnar. Wydawnictwo Naukowe PWN, Warszawa 2012.

Walesiak M.: Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R. UE we Wrocławiu, Wrocław 2011.

Walesiak M., Dudek A.: clusterSim: Searching for optimal clustering procedure for a data set, R package ver. 0.43-4, 2014, http://cran.r-project.org/web/packages/clusterSim/index.html.

Warnes G.R., Bolker B., Bonebakker L., Gentleman R., Huber W., Liaw A., Lumley T., Maechler M., Magnusson A., Moeller S., Schwartz M., Venables B.: gplots: Various R programming tools for plotting data, R package ver. 2.13.0, 2014, http://cran.r-project.org/web/packages/gplots/index.html.

Wei T.: corrplot: Visualization of a correlation matrix, R package ver. 0.73, 2013, http://cran.r-roject.org/web/packages/corrplot/index.html.

Wickham H.: ggplot2: elegant graphics for data analysis. Springer, New York 2009.

SUPPORTING ANALYSES BASED ON VARIABLES MEASURED ON VARIOUS SCALES BY VISUALIZATION TECHNIQUES

Summary

The article has a methodological and applicative objective associated with the graphical presentation of datasets, similarities of objects and results of multivariate analyses. The paper gives an outline of procedures useful in situations in which variables are measured on various measurement scales. The research approach is based on literature studies and analyses of secondary data relating to purchases on the Internet. The given examples show that data visualization methods can provide a valuable support in conducting social research in case of possessing a dataset containing variables measured on various scales.