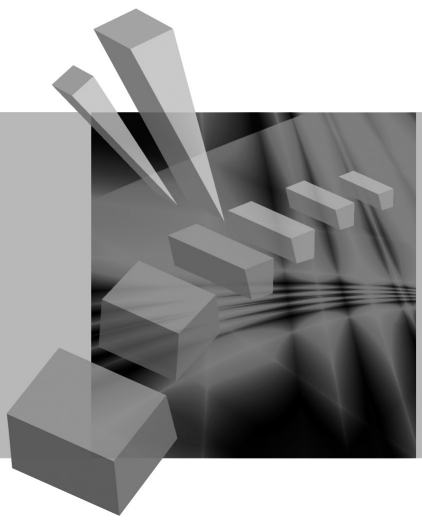
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
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

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana
Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Justyna Brzezińska
Uniwersytet Ekonomiczny w Katowicach

METODY WIZUALIZACJI DANYCH JAKOŚCIOWYCH W PROGRAMIE R

Streszczenie: W artykule zaprezentowane zostaną graficzne metody analizy danych jakościowych. Metody wizualizacji są dobrze rozwinięte w przypadku danych metrycznych; jeśli chodzi o dane niemetryczne, są one dopiero w fazie rozwoju. Wiele z nich jest wyspecjalizowanym narzędziem służącym do analizy tablic kontyngencji dowolnego typu, jak np. wykres mozaikowy, wykres *double-decker*, wykres sitkowy. Metodom wizualizacji danych niemetrycznych towarzyszą znane już w wielowymiarowej analizie statystycznej metody analizy danych, takie jak analiza logarytmiczno-liniowa oraz analiza korespondencji, które w szczegółowy sposób analizują strukturę badanego zjawiska. Techniki wizualizacji danych dostępne są w pakiecie `vcd` oraz `vcdExtra` w programie **R**.

Słowa kluczowe: analiza danych niemetrycznych, wizualizacja danych niemetrycznych, tablice kontyngencji, modele logarytmiczno-liniowe, analiza korespondencji.

1. Wstęp

W badaniach ekonomiczno-społecznych szczególną rolę odgrywają zmienne niemetryczne, tj. takie, które mierzone są na słabych skalach pomiaru. Zmienne te przedstawiane są zazwyczaj w dwu- lub wielowymiarowych tablicach kontyngencji, a miernikami badania zależności są współczynniki: Yule'a, Pearsona i Cramera, Czuprowa. Współczynniki te wykorzystywane są zazwyczaj w sytuacjach, gdy tablica kontyngencji jest dwuwymiarowa. W sytuacjach, gdy mamy do czynienia z tablicą wielowymiarową, skuteczną metodą analizy tego typu danych jest analiza logarytmiczno-liniowa, analiza korespondencji lub też, w przypadku danych brakujących lub nieobserwowalnych, analiza klas ukrytych. Dodatkową zaletą wymienionych metod jest możliwość zaprezentowania ich struktury w postaci graficznej za pomocą odpowiednich wykresów.

Istnieje wiele metod graficznych przeznaczonych do wizualizacji danych jakościowych, jednak w niniejszej pracy zaprezentowane zostaną jedynie niektóre z nich, takie jak: wykres *fourfold* (*fourfold display*), sitkowy (*sieve plot*), mozaikowy (*mosaic plot*) oraz wykres asocjacji (*association plot*). Każdy z nich związany jest z graficzną prezentacją odchyleń liczebności empirycznych od teoretycznych występują-

cych w tablicy kontyngencji (im odchylenia są mniejsze, tym model jest lepiej dopasowany do danych). W przypadku wielu zmiennych metody wizualizacji ułatwiają wybór modelu najlepiej dopasowanego do danych, a także pozwalają na szczegółową analizę związku pomiędzy zmiennymi.

W niniejszym artykule metody wizualizacji danych jakościowych zaprezentowane zostaną kolejno dla tablic dwu- oraz wielowymiarowych na przykładzie danych dotyczących wymiaru czasu pracy oraz struktury bezrobocia w Polsce w 2011 r. Dane wykorzystane w badaniu pochodzą z Banku Danych Lokalnych Głównego Urzędu Statystycznego (www.stat.gov.pl). Metody wizualizacji dostępne są w programie **R** pakiecie `vcd` oraz `vcdExtra`.

2. Metody wizualizacji danych jakościowych w programie R

2.1. Tablice kontyngencji 2 × 2

Jednym z prostszych wykresów, który przeznaczony jest dla tablic kontyngencji o wymiarach 2×2, jest wykres *fourfold* (*fourfold display*), w którym liczebność n_{hj} dla każdej komórki przedstawione jest w postaci ćwiartki koła, którego promień jest proporcjonalny do $\sqrt{n_{hj}}$. Wykres ten jest analogiczny do wykresu kołowego, jednak różnicą jest kąt koła pomiędzy wycinkiem koła, który w wykresie kołowym jest zmienny, a w wykresie *fourfold* stały (90°), a także promień koła, który na wykresie kołowym jest stały, a na wykresie *fourfold* zmienny [Fienberg 1975; Friendly 1994]. Na wykresie w postaci łuków wewnętrznych i zewnętrznych przedstawione są przedziały ufności ilorazu szans θ na ustalonym poziomie ufności $\gamma = 0,95$ (wartości γ mogą być zmieniane). W rogu każdej ćwiartki wykresu przedstawione są liczebności każdej z komórki tablicy kontyngencji. Wykres ten stanowi graficzną prezentację hipotezy o niezależności postaci:

$H_0 : \theta = 1$ (zmienne są niezależne),

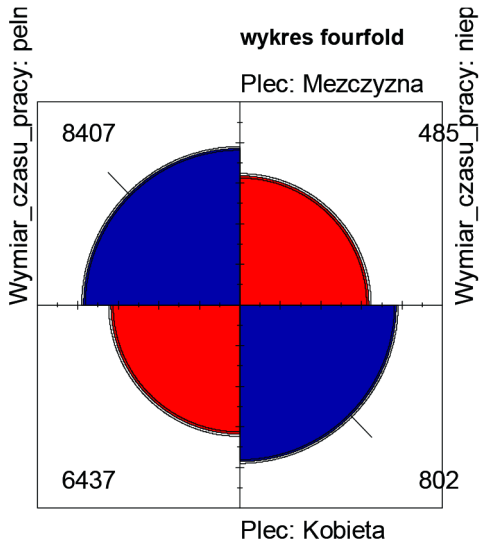
$H_1 : \theta \neq 1$ (zmienne są zależne).

Jeśli liczebności empiryczne są większe od teoretycznych, wówczas dana ćwiartka koła oznaczona jest kolorem niebieskim. W przeciwnym wypadku, gdy liczebności empiryczne są mniejsze od teoretycznych, ćwiartka koła oznaczona jest kolorem czerwonym.

Wykres *fourfold* uzyskany dzięki funkcji `fourfold{vcd}` zaprezentowano z wykorzystaniem danych z Głównego Urzędu Statystycznego dotyczących wymiaru czasu pracy względem płci w 2011 r. dla próby liczącej 16 131 osób.

Kolory[1] na rys. 1 oznaczają znak różnic między liczebnościami empirycznymi a teoretycznymi, a wartość ilorazu szans $\theta = 2,16$ oznacza, że prawdopodobieństwo

[1] Znak różnicy między liczebnościami oznaczono na rys. 1 kolorem czerwonym dla różnic ujemnych, a niebieskim dla różnic dodatnich.



Rys. 1. Wykres *fourfold* dla tablicy kontyngencji 2×2

Źródło: opracowanie własne w **R** na podstawie danych z Banku Danych Lokalnych Głównego Urzędu Statystycznego (www.stat.gov.pl).

wystąpienia sukcesu w pierwszym wierszu jest ponaddwukrotnie wyższe niż w drugim. Dodatni znak współczynnika oznacza, że zależność między zmiennymi jest zgodna co do kierunku.

2.2. Tablice kontyngencji $H \times J$

W przypadku analizy dwuwymiarowych tablic kontyngencji $H \times J$ popularnymi wykresami są wykres sitkowy i mozaikowy. Niezależność zmiennych przedstawiona jest poprzez wyrażenie liczebności oczekiwanych jako iloczynu liczebności brzegowych wierszy i kolumn podzielonych przez całkowitą liczebność tablicy. Riedwyl i Schüpbach [1983; 1994] wprowadzili do literatury pojęcie wykresu sitkowego (*sieve diagram*), nazwanego także wykresem parkietowym (*parquet diagram*). Na wykresie tym powierzchnia każdego prostokąta jest proporcjonalna do liczebności oczekiwanych $\hat{m}_{hj}$, przy czym liczebność empiryczna odpowiada liczbie kwadratów w danym prostokącie [Friendly 2000]. Szerokość każdego prostokąta jest proporcjonalna do liczebności brzegowych kolumn $n_{\bullet j}$, a wysokość do liczebności brzegowych wierszy $n_{h\bullet}$. Odchylenia liczebności empirycznych od teoretycznych ($n_{hj} - \hat{m}_{hj}$) oznaczone są w postaci kolorów. Jeśli różnica ta jest ujemna, wówczas linia tworząca kwadraty w odpowiednim prostokącie jest czerwoną linią ciągłą. Jeśli

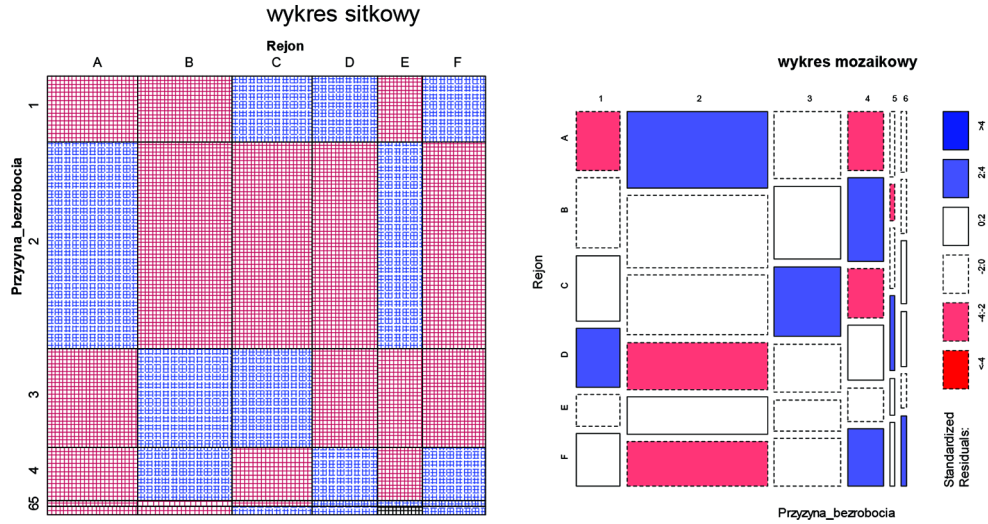
różnica ta jest dodatnia, wówczas linia w danym prostokącie jest przerywaną niebieską. Niezależność pomiędzy zmiennymi występuje wówczas, gdy zagęszczenie i struktura kwadratów jest jednorodna. W przypadku niejednorodności można sądzić, że zmienne są zależne [Friendly 2012].

Wykres mozaikowy został wprowadzony do literatury przez Hartigana i Kleinera [1981; 1984] oraz Theusa [1997] i stanowi metodę graficznej prezentacji wyników modeli w wielowymiarowych tablicach kontyngencji. Pomimo że jest to bardzo istotny krok w analizie danych niemetrycznych, metoda ta nie jest popularna, a jej rozwój przypada na koniec XX i początek XXI wieku. Wykres mozaikowy jest graficzną prezentacją liczebności tablicy kontyngencji. Dzięki niemu możliwa jest także graficzna ocena modelu. Wykresy mozaikowe mają charakterystyczny kształt zależny od postaci równania modelu i zawartych w nich parametrów odpowiadającym badanym zmiennym. Ten kształt odzwierciedla strukturę modelu, zależną od występowania lub braku w równaniu modelu danego współczynnika.

Wykresy mozaikowe składają się z prostokątnych płytek (*tile, bin, box, rectangle*), których pole jest proporcjonalne do liczebności empirycznej n_{hj}, szerokość proporcjonalna jest do liczebności brzegowej $n_{h\bullet}$, a wysokość do proporcji $\frac{n_{hj}}{n_{h\bullet}}$. Budowa tego wykresu oparta jest na standaryzowanych resztach Pearsona, zdefiniowanych jako: $d_{hj} = \frac{n_{hj} - \hat{m}_{hj}}{\sqrt{\hat{m}_{hj}}}$. Jeśli reszta jest dodatnia, dany prostokąt oznaczony jest kolorem niebieskim, jeśli ujemna – kolorem czerwonym. Przedziały, w których znajdują się reszty, oznaczone są coraz ciemniejszym kolorem w miarę wzrostu wartości d_{hj} ($\left|d_{hj}\right| > 0, 2, 4, ...$).

Do graficznej prezentacji wykresu sitkowego oraz mozaikowego wykorzystano zbiór danych z Głównego Urzędu Statystycznego dotyczący przyczyn bezrobocia w różnych rejonach Polski w 2011 r. dla 13 484 osób. Zbudowano dwuwymiarową tablicę kontyngencji o wymiarach 6×6 dla zmiennych: „Przyczyna bezrobocia" (1. choroba lub niepełnosprawność, 2. emerytura, 3. nauka i uzupełnienie kwalifikacji, 4. obowiązki rodzinne, 5. wyczerpane wszystkie możliwości poszukiwania pracy, 6. przekonanie o niemożliwości znalezienia pracy) oraz „Rejon" (A. centralny, B. południowy, C. wschodni, D. północno-zachodni, E. południowo-zachodni, F. północny). Ze względu na długie nazwy kategorii na wykresie zarówno sitkowym, jak i mozaikowym wykorzystano jedynie symbole zamiast pełnych nazw kategorii. Wykres sitkowy i mozaikowy dostępny jest w programie **R** dzięki funkcjom: `mosaic{vcd}` oraz `sieve{vcd}`.

Wykres sitkowy i mozaikowy zaprezentowane są na rys. 2.

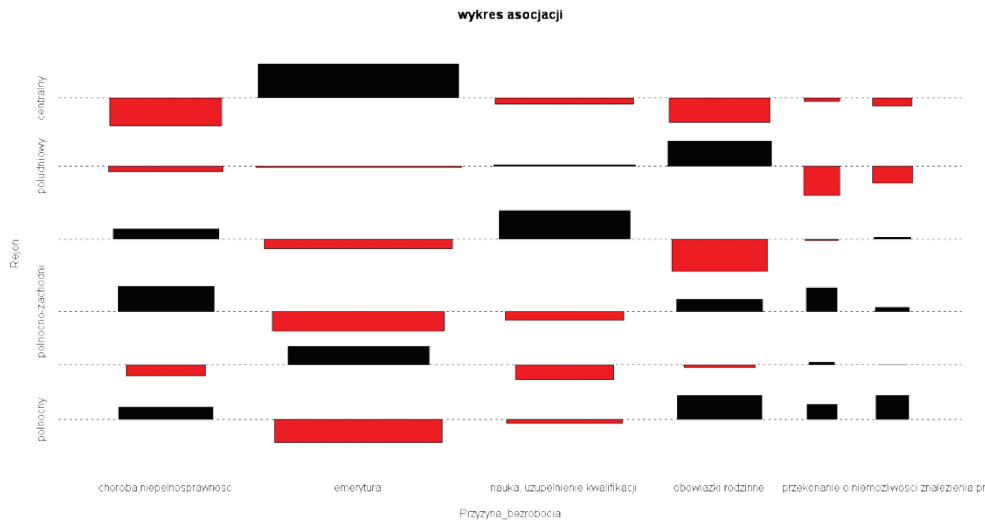


Rys. 2. Wykres sitkowy i mozaikowy dla dwuwymiarowej tablicy kontyngencji

Źródło: opracowanie własne w **R** na podstawie danych z Banku Danych Lokalnych Głównego Urzędu Statystycznego (www.stat.gov.pl).

Z analizy wykresu sitkowego zaobserwować można, że częściej przyczyną bezrobocia jest choroba i niepełnosprawność w rejonie wschodnim, północno-zachodnim oraz północnym niż w rejonie centralnym i południowym. Oznacza to, że osoby z tej grupy mają najmniejszy wpływ na odrzucenie hipotezy zerowej o niezależności zmiennych. Podobnie interpretować można pozostałe przyczyny bezrobocia w danym rejonie Polski. Im większe zagęszczenie w prostokącie, tym większe występują odchylenia pomiędzy liczebnościami empirycznymi a teoretycznymi, a tym samym większe odchylenia od niezależności. Wykres mozaikowy stosowany jest do badania niezależności w sposób graficzny; im. zacienienie jest mocniejsze, tym silniejsza niezależność. Komórki puste w całym obszarze świadczą o niezależności.

Wykres asocjacji jest kolejnym wykresem (rys. 3) wizualizacji zmiennych niemetrycznych w wielowymiarowych tablicach kontyngencji, na którym prostokąty są proporcjonalne do liczebności teoretycznych m_{hj}. Odchylenia liczebności empirycznych od teoretycznych zaznaczone są kolorami. Jeśli różnica ta jest ujemna, wówczas prostokąt jest czerwony i znajduje się poniżej linii, jeśli różnica ta jest dodatnia, wówczas prostokąt jest czarny i usytuowany jest powyżej linii. Wysokość prostokąta jest proporcjonalna do standaryzowanej reszty Pearsona d_{hj}, a szerokość do $\sqrt{m_{hj}}$. Wykres asocjacji w programie **R** dostępny jest dzięki funkcji `assocplot {graphics}`.



Rys. 3. Wykres asocjacji dla dwuwymiarowej tablicy kontyngencji

Źródło: opracowanie własne w **R** na podstawie danych z Banku Danych Lokalnych Głównego Urzędu Statystycznego (www.stat.gov.pl).

Interpretacja wykresu asocjacji jest tutaj trudna, gdyż struktura odchyleń jest zmienna i nie można zaobserwować wzrostu czy też spadku odchyleń dla którejś z badanych kategorii. Kolory i wielkość prostokątów mówią jednak o znaku odchyleń w każdej komórce, a ich wielkość o ich rozmiarze.

2.3. Wielowymiarowe tablice kontyngencji

W przypadku wielowymiarowych tablic kontyngencji wykresy mozaikowe służą najczęściej do zaprezentowania struktury danych i rodzaju powiązań między zmiennymi, ale również do oceny jakości dopasowania danego modelu do danych w sposób graficzny.

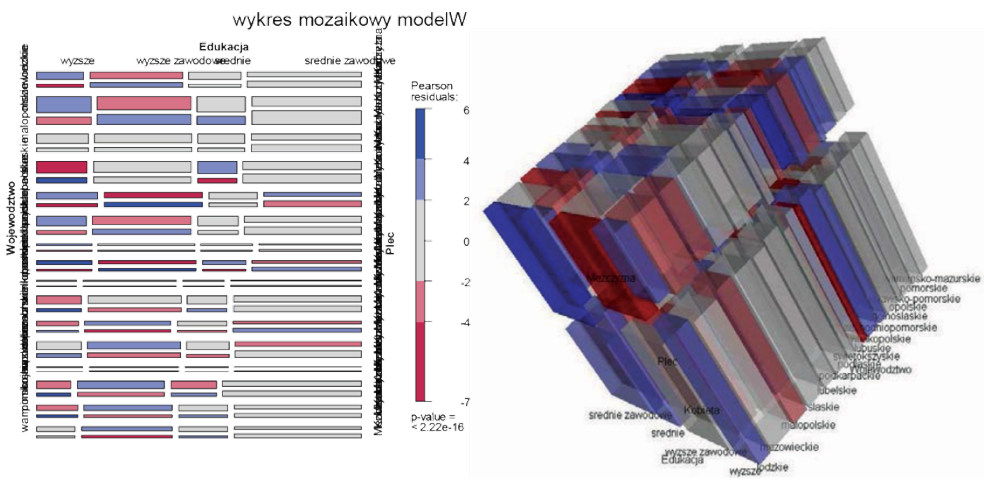
Zbudowano trójwymiarową tablicę kontyngencji $H \times J \times K$ dla zmiennych: „Województwo", „Wykształcenie bezrobotnego" oraz „Płeć" dla próby liczącej 1 436 814 osoby. Zbudowano wszystkie modele logarytmiczno-liniowe z trzema zmiennymi, dla których wyznaczono współczynniki: chi-kwadrat, iloraz wiarygodności oraz kryteria informacyjne AIC oraz BIC (tab. 1).

Modelem najlepiej dopasowanym do danych jest model zależności homogenicznej $[WE][WP][EP]$, dla którego współczynniki te osiągają wartość najmniejszą i oznaczają najmniejsze odchylenia liczebności empirycznych od teoretycznych. Wykres mozaikowy w przestrzeni dwu- i trójwymiarowej dostępny jest w programie **R** dzięki funkcjom: `mosaic{vcd}` oraz `mosaic3d{vcdExtra}`. Dla tego modelu zaprezentowano wykres mozaikowy w przestrzeni dwu- i trójwymiarowej (rys. 4).

Tabela 1. Kryteria dopasowania modeli do danych

Symbol	χ^2	G^2	*df*	*p-value*	*AIC*	*BIC*
$[W][P][E]$	77 280,71	79 283,84	108	0	79 067,84	77 752,62
$[EP][W]$	18 239,83	18 314,23	105	0	18 104,23	16 825,55
$[WE][P]$	66 999,68	67 562,55	63	0	67 436,55	66 669,34
$[WP][E]$	75 805,59	76 031,78	93	0,014	75 845,78	74 713,23
$[WP][EP]$	14 948,51	15 062,17	90	0,291	14 882,17	13 786,16
$[WE][EP]$	6 593,30	6 592,94	60	0,014	6 472,94	5742,26
$[WE][WP]$	63 838,36	64 310,49	48	0	64 214,49	63 629,95
$[WE][WP][EP]$	1 005,21	1 005,52	45	0	915,52	367,51
$[WPE]$	0	0	0	1	0	0

Źródło: opracowanie własne w programie **R**.



Rys. 4. Wykres mozaikowy w przestrzeni dwu- i trójwymiarowej dla modelu zależności homogenicznej $[WE][WP][EP]$

Źródło: opracowanie własne w **R** na podstawie danych z Banku Danych Lokalnych Głównego Urzędu Statystycznego (www.stat.gov.pl).

Kolory odpowiadające kolejnym komórkom tablicy kontyngencji świadczą o niewielkich odchyleniach liczebności empirycznych od teoretycznych, zatem model zależności homogenicznej $[WE][WP][EP]$ można uznać za model dobrze dopasowany do danych. W modelu tym ilorazy szans wyznaczone dla każdej pary zmiennych są identyczne dla każdej wartości trzeciej zmiennej. Jest to model o dużej zło-

żoności, gdyż zawiera trzy interakcje pomiędzy wszystkimi zmiennymi, jednak ze względu na to, iż celem artykułu jest jedynie prezentacja metod wizualizacji, struktura modelu i interpretacja jego parametrów zostaną pominięte.

3. Podsumowanie

Zaawansowane programy komputerowe w ostatnich latach przyczyniły się do wzrostu zainteresowania metodami analizy danych jakościowych, które przez długi czas pozostawały w cieniu metod przeznaczonych dla danych ilościowych. Dane jakościowe, mierzone na słabych skalach pomiaru (nominalna lub porządkowa), zapisywane są zazwyczaj w formie tablic kontyngencji (dwu- lub wielowymiarowych). Wizualizacja tego rodzaju danych będąca tematem niniejszego artykułu daje szerokie możliwości określenia rodzaju zależności między zmiennymi, przedstawiając tym samym w szczegółowy sposób strukturę badanego zjawiska. Jest to szczególnie przydatne w sytuacjach, gdy analizie poddanych jest kilka zmiennych jednocześnie. Metody wizualizacji danych niemetrycznych zaprezentowane w niniejszym artykule z powodzeniem wykorzystywane mogą być jako uzupełnienie klasycznej analizy danych, jak np. analiza zależności, analiza korespondencji czy też analiza logarytmiczno-liniowa. Dzięki odpowiednim wykresom, jak np. wykres sitkowy, mozaikowy czy też wykres asocjacji, możliwe jest przedstawienie odchyleń liczebności empirycznych od teoretycznych w danej tablicy kontyngencji w sposób graficzny, a co za tym idzie – ocena jakości dopasowania. Narzędzia wizualizacyjne są szczególnie przydatne w sytuacjach, gdy formalny model jest skomplikowany, a interpretacja jego parametrów trudna. Pakiety `vcd`, `graphics` oraz `vcdExtra` dostępne w programie **R** pozwalają na graficzną prezentację zmiennych zapisanych w postaci tablic kontyngencji dowolnego wymiaru. W niniejszym artykule metody wizualizacji danych jakościowych zostały wykorzystane do zaprezentowania struktury bezrobocia w Polsce w 2011 r.

Literatura

Friendly M. (1994), *Mosaic displays for multi-way contingency tables*, "Journals of the American Statistical Association", 49, 153-160.

Friendly M. (1995), *Conceptual and visual models for categorical data*, "The Amercian Statistician", 49, 153-160.

Friendly M. (2000), *Visualizing Categorical Data*, SAS Institute.

Friendly M. (2012), *Visualizing Data with SAS and R*, York University Short Course, www:datavis.ca/courses/VCD.

Fienberg S.E. (1975), *Perspective Canada as a social report*, Social Indicators Research, 2, 153-174.

Hartigan J.A., Kleiner B. (1981), *Mosaics for Contingency Tables*, [w:] *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, ed. W.F. Eddy, Springer, New York, 268-273.

Hartigan J.A., Kleiner B. (1984), *A mosaic of television ratings*, "The American Statistician", 38, 32-35.
Riedwyl H., Schüpbach M. (1983), *Siebdiagramme: Graphische darstellung von kontingenztafeln*, *Technical Report*, 12, Institute for Mathematical Statistics, University of Bern, Bern, Switzerlad.
Riedwyl H., Schüpbach M. (1994), *Parquet Diagram to Plot Contingency Tables, In Faulbaum, F., editor, Softstat `93*: *Advanced in Statistical Software*, Gustav Fischer, New York, 293-299.
Theus M. (1997), *Visualization of categorical data*, *Advanced in Statistical Software*, Lucius & Lucius, 6, 47-55.

VISUALIZING CATEGORICAL DATA IN R

Summary: This paper presents the use of graphical methods for the analysis of multi-way contingency table. Graphical methods for categorical data are well known and fully developed, however, visualizing categorical data is only now being developed. Many of these are specialized for particular types of tables and most are not readily available in standard software, and they are not widely used. In this paper we illustrate the use of mosaic displays and other graphical methods for the analysis of several multi-way contingency tables e.g. sieve plot, double-decker plot, fourfold plot. Second, we introduce several extensions of mosaic displays designed to integrate graphical methods for categorical data with those used for categorical data.

Keywords: categorical variable, visualizing categorical data, cross-table, log-linear analysis, correspondence analysis.