

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

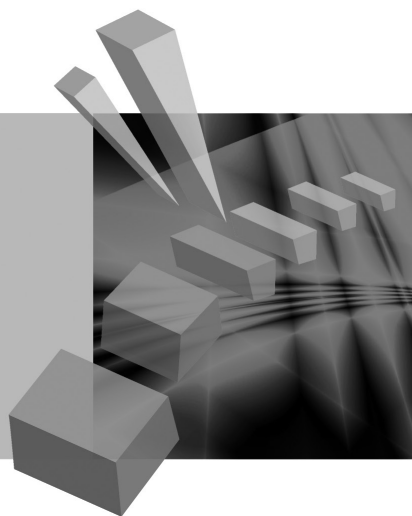
RESEARCH PAPERS

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

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2013

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

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: Sejm VI kadencji – maszynka do głosowania	11
Barbara Pawelek, Adam Sagan: Zmienne ukryte w modelach ekonomicznych – respecyfikacja modelu Kleina I	19
Jan Paradysz: Nowe możliwości badania koniunktury na rynku pracy	29
Krzysztof Najman: Samouczące się sieci GNG w grupowaniu dynamicznym zbiorów o wysokim wymiarze	41
Kamila Migdał-Najman: Zastosowanie jednowymiarowej sieci SOM do wyboru cech zmiennych w grupowaniu dynamicznym	48
Aleksandra Matuszewska-Janica, Dorota Witkowska: Zróżnicowanie płac ze względu na płeć: zastosowanie drzew klasyfikacyjnych	58
Iwona Foryś, Ewa Putek-Szeląg: Przestrzenna klasyfikacja gmin ze względu na sprzedaż użytków gruntowych zbywanych przez ANR w województwie zachodniopomorskim	67
Joanna Banaś, Małgorzata Machowska-Szewczyk: Klasyfikacja internetowych rachunków bankowych z uwzględnieniem zmiennych symbolicznych.....	77
Marta Jaročka: Wpływ metody doboru cech diagnostycznych na wynik porządkowania liniowego na przykładzie rankingu polskich uczelni	85
Anna Zamojska: Badanie zgodności rankingów wyznaczonych według różnych wskaźników efektywności zarządzania portfelem na przykładzie funduszy inwestycyjnych.....	95
Dorota Rozmus: Porównanie dokładności taksonomicznej metody propagacji podobieństwa oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	106
Ewa Wędrowska: Wrażliwość miar dywergencji jako mierników niepodobieństwa struktur.....	115
Katarzyna Wójcik, Janusz Tuchowski: Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich ...	124
Małgorzata Misztal: Ocena wpływu wybranych metod imputacji na wyniki klasyfikacji obiektów w modelach drzew klasyfikacyjnych.....	135
Anna Czapkiewicz, Beata Basiura: Badanie wpływu wyboru współczynnika zależności na grupowanie szeregów czasowych	146
Tomasz Szubert: Czynniki różnicujące poziom zadowolenia z życia oraz wartości życiowe osób sprawnych i niepełnosprawnych w świetle badań „Diagnozy społecznej”	154

Marcin Szymkowiak: Konstrukcja estymatorów kalibracyjnych wartości globalnej dla różnych funkcji odległości	164
Wojciech Roszka: Szacowanie łącznych charakterystyk cech nieobserwowanych łącznie	174
Justyna Brzezińska: Metody wizualizacji danych jakościowych w programie R	182
Agata Sielska: Regionalne zróżnicowanie potencjału konkurencyjnego polskich gospodarstw rolnych w województwach po akcesji do Unii Europejskiej	191
Mariusz Kubus: Liniowy model prawdopodobieństwa z regularyzacją jako metoda doboru zmiennych	201
Beata Basiura: Metoda Warda w zastosowaniu klasyfikacji województw Polski z różnymi miarami odległości	209
Katarzyna Wardzińska: Wykorzystanie metody obwiedni danych w procesie klasyfikacji przedsiębiorstw	217
Katarzyna Dębowska: Modelowanie upadłości przedsiębiorstw oparte na próbach niezbilansowanych	226
Danuta Tarka: Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska ..	235
Artur Czech: Zastosowanie wybranych metod doboru zmiennych diagnostycznych w badaniach konsumpcji w ujęciu pośrednim	246
Beata Bal-Domańska: Ocena relacji zachodzących między inteligentnym rozwojem a spójnością ekonomiczną w wymiarze regionalnym z wykorzystaniem modeli panelowych	255
Mariola Chrzanowska: <i>Ordinary kriging</i> i <i>inverse distance weighting</i> jako metody szacowania cen nieruchomości na przykładzie warszawskiego rynku	264
Adam Depta: Zastosowanie analizy wariancji w badaniu jakości życia na podstawie kwestionariusza SF-36v2	272
Maciej Beręsewicz, Tomasz Klimanek: Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach cen mieszkań	281
Karolina Paradysz: Benchmarkowa analiza estymacji dla małych obszarów na lokalnych rynkach pracy	291
Anna Gryko-Nikitin: Dobór parametrów w równoległych algorytmach genetycznych dla problemu plecakowego	301
Tomasz Ząbkowski, Piotr Jałowicki: Zastosowanie reguł asocjacyjnych do analizy danych ankietowych w wybranych obszarach logistyki przedsiębiorstw przetwórstwa rolno-spożywczego	311
Agnieszka Przedborska, Małgorzata Misztal: Zastosowanie metod statystyki wielowymiarowej do oceny wydolności stawów kolanowych u pacjentów z chorobą zwyrodnieniową leczonych operacyjnie	321
Dorota Perło: Rozwój zrównoważony w wymiarze gospodarczym, społecznym i środowiskowym – analiza przestrzenna	331

Ewa Putek-Szeląg, Urszula Gieraltowska, Analiza i diagnoza wielkości produkcji energii odnawialnej w Polsce na tle krajów Unii Europejskiej..	342
--	-----

Summaries

Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: VIth-term Sejm – a voting machine	18
Barbara Pawelek, Adam Sagan: Latent variables in econometric models – respecification of Klein I model	28
Jan Paradysz: New possibilities for studying the situation on the labour market	40
Krzysztof Najman: Self-learning neural network of GNG type in the dynamic clustering of high-dimensional data.....	47
Kamila Migdał-Najman: Applying the one-dimensional SOM network to select variables in dynamic clustering	57
Aleksandra Matuszewska-Janica, Dorota Witkowska: Gender wage gap: application of classification trees.....	66
Iwona Foryś, Ewa Putek-Szeląg: Spatial classification of communes by usable land traded by the APA in the Zachodniopomorskie voivodeship...	76
Joanna Banaś, Małgorzata Machowska-Szewczyk: Classification of Internet banking accounts including symbolic variables	84
Marta Jarocka: The impact of the method of the selection of diagnostic variables on the result of linear ordering on the example of ranking of universities in Poland.....	94
Anna Zamojska: Empirical analysis of the consistency of mutual fund ranking for different portfolio performance measures.....	105
Dorota Rozmus: Comparison of accuracy of affinity propagation clustering and cluster ensembles based on bagging idea.....	114
Ewa Wędrowska: Sensitivity of divergence measures as structure dissimilarity measurements	123
Katarzyna Wójcik, Janusz Tuchowski: Machine translation impact on the results of the sentiment analysis	134
Małgorzata Misztal: Assessment of the influence of selected imputation methods on the results of object classification using classification trees ...	145
Anna Czapkiewicz, Beata Basiura: Simulation study of the selection of coefficient depending on the clustering time series.....	153
Tomasz Szubert: Factors differentiating the level of satisfaction with life and the life's values of people with and without disabilities in the light of the "Social Diagnosis" survey	162
Marcin Szymkowiak: Construction of calibration estimators of totals for different distance measures	173

Wojciech Roszka: Joint characteristics' estimation of variables not jointly observed.....	181
Justyna Brzezińska: Visualizing categorical data in \mathbf{R}	190
Agata Sielska: Regional diversity of competitiveness potential of Polish farms after the accession to the European Union	200
Mariusz Kubus: Regularized linear probability model as a filter	208
Beata Basiura: The Ward method in the application for classification of Polish voivodeships with different distances.....	216
Katarzyna Wardzińska: Application of Data Envelopment Analysis in company classification process.....	225
Katarzyna Dębowska: Modeling corporate bankruptcy based on unbalanced samples	234
Danuta Tarka: Influence of the features selection method on the results of objects classification using environmental data.....	245
Artur Czech: Application of chosen methods for the selection of diagnostic variables in indirect consumption research.....	254
Beata Bal-Domańska: Assessment of relations occurring between smart growth and economic cohesion in regional dimension using panel models	263
Mariola Chrzanowska: Ordinary kriging and inverse distance weighting as methods of estimating prices based on Warsaw real estate market	271
Adam Depta: Application of analysis of variance in the study of the quality of life based on questionnaire SF-36v2	280
Maciej Beręsewicz, Tomasz Klimanek: Using indirect estimation with spatial autocorrelation in dwelling price surveys.....	290
Karolina Paradysz: Benchmark analysis of small area estimation on local labor markets	300
Anna Gryko-Nikitin: Selection of various parameters of parallel evolutionary algorithm for knapsack problems	310
Tomasz Ząbkowski, Piotr Jałowiecki: Application of association rules for the survey of data analysis in the selected areas of logistics in food processing companies	320
Agnieszka Przedborska, Małgorzata Misztal: Using multivariate statistical methods to assess the capacity of the knee joint among the patients treated surgically for osteoarthritis	330
Dorota Perło: Sustainable development in the economic, social and environmental dimensions – spatial analysis.....	341
Ewa Putek-Szeląg, Urszula Gieraltowska: Analysis and diagnosis of the volume of renewable energy production in Poland compared to EU countries	352

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 odchyłeń 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_{ij} dla każdej komórki przedstawione jest w postaci ćwiartki koła, którego promień jest proporcjonalny do $\sqrt{n_{ij}}$. 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:

$$H_0 : \theta = 1 \text{ (zmiennie są niezależne),}$$

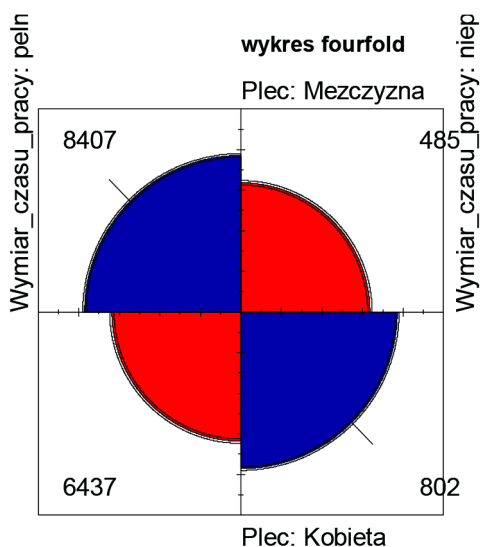
$$H_1 : \theta \neq 1 \text{ (zmiennie są zależne).}$$

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.

Kolory¹ 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 siatkowy 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 siatkowego (*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 $\frac{n_{hj}}{n_{h\bullet}}$.

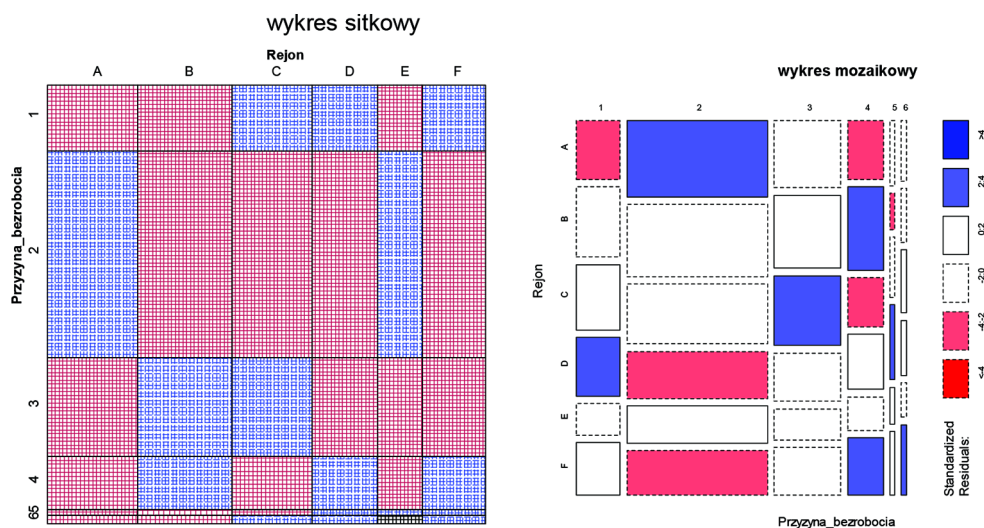
Budowa tego wykresu oparta jest na standaryzowanych resztach Pearsona, zdefiniowanych jako:

$$d_{hj} = \frac{n_{hj} - \hat{m}_{hj}}{\sqrt{\hat{m}_{hj}}}$$

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} ($|d_{hj}| > 0, 2, 4, \dots$).

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.

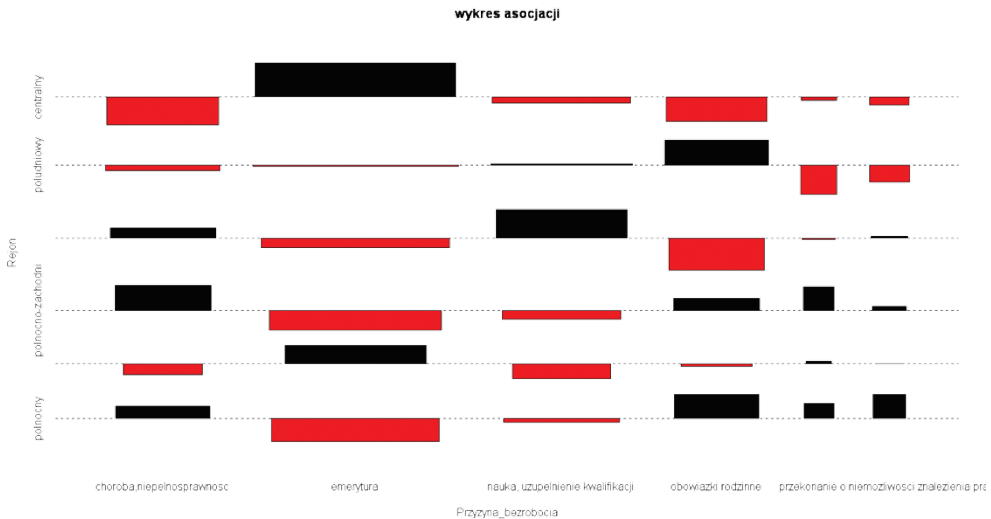


Rys. 2. Wykres sitkowy i mozaikowy dla dwuwymiarowej tablicy kontyngencji

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 zaciemnienie 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 odchyłeń jest zmienna i nie można zaobserwować wzrostu czy też spadku odchyłeń dla którejś z badanych kategorii. Kolory i wielkość prostokątów mówią jednak o znaku odchyłeń 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).

zoności, 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 siatkowy, mozaikowy czy też wykres asocjacji, możliwe jest przedstawienie odchyżeń 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 American 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, Switzerland.
- 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.