PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
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

Wstęp 9

Eugeniusz Gatnar, Balance of payments statistics and external competitiveness of Poland 15

Andrzej Sokołowski, Magdalena Czaja, Efektywność metody *k*-średnich w zależności od separowalności grup 23

Barbara Pawełek, Józef Pociecha, Adam Sagan, Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw 30

Elżbieta Gołata, Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011 39

Aleksandra Łuczak, Feliks Wysocki, Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów 49

Marek Walesiak, Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej 60

Paweł Lula, Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu *RAKE* 69

Mariusz Kubus, Propozycja modyfikacji metody złagodzonego LASSO 77

Andrzej Bąk, Tomasz Bartłomowicz, Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie `DiscreteChoice` programu R 85

Justyna Brzezińska, Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012 95

Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Zastosowanie pakietu `dcMNM` programu R w badaniach preferencji konsumentów wódki 104

Barbara Batóg, Jacek Batóg, Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010 113

Małgorzata Markowska, Danuta Strahl, Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego 121

Kamila Migdał-Najman, Krzysztof Najman, Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 131

Kamila Migdał-Najman, Krzysztof Najman, Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 139

Beata Basiura, Anna Czapkiewicz, Badanie jakości klasyfikacji szeregów czasowych 148

Michał Trzęsiok, Wybrane metody identyfikacji obserwacji oddalonych 157

Grażyna Dehnel, Tomasz Klimanek, Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej ... 167
Michał Bernard Pietrzak, Justyna Wilk, Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji ... 177
Maciej Beręsewicz, Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena ... 186
Marcin Szymkowiak, Tomasz Józefowski, Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji ... 195
Marcin Pełka, Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym ... 202
Małgorzata Machowska-Szewczyk, Ocena klas w rozmytej klasyfikacji obiektów symbolicznych ... 210
Justyna Wilk, Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych ... 220
Andrzej Dudek, Metody analizy skupień w klasyfikacji markerów map Google ... 229
Ewa Roszkowska, Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW 237
Marcin Szymkowiak, Marek Witkowski, Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych ... 248
Bartłomiej Jefmański, Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R ... 257
Karolina Bartos, Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych ... 266
Joanna Trzęsiok, Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych ... 275
Beata Bal-Domańska, Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja ... 285
Beata Bieszk-Stolorz, Iwona Markowicz, Wpływ zasiłku na proces poszukiwania pracy ... 294
Marta Dziechciarz-Duda, Klaudia Przybysz, Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni ... 303
Tomasz Klimanek, Problem pomiaru procesu dezagraryzacji wsi polskiej w świetle wielowymiarowych metod statystycznych ... 313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Wybrane metody analizy danych wzdłużnych ... 321
Artur Zaborski, Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych ... 330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy ... 338

Katarzyna Wawrzyniak, Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego 346

Summaries

Eugeniusz Gatnar, Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski 22
Andrzej Sokołowski, Magdalena Czaja, Cluster separability and the effectiveness of *k*-means method 29
Barbara Pawełek, Józef Pociecha, Adam Sagan, Multisectoral analysis of latent transitions in bankruptcy prediction models..... 38
Elżbieta Gołata, Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011 48
Aleksandra Łuczak, Feliks Wysocki, Determination of weights for features in problems of linear ordering of objects 59
Marek Walesiak, Reinforcing measurement scale for ordinal data in multivariate statistical analysis 68
Paweł Lula, Automatic identification of keywords and keyphrases in documents written in Polish..... 76
Mariusz Kubus, The proposition of modification of the relaxed LASSO method..... 84
Andrzej Bąk, Tomasz Bartłomowicz, Microeconomic multinomial logit models and their implementation in the `DiscreteChoice` R package. 94
Justyna Brzezińska, The analysis of unemployment data in Poland in 2004--2012 with application of log-linear models 103
Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Application of the MMLM package of R software for vodka consumers preference analysis..... 112
Barbara Batóg, Jacek Batóg, Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity 120
Małgorzata Markowska, Danuta Strahl, Classification of the European regional space in terms of smart growth pillars using the reference limit system..... 130
Kamila Migdał Najman, Krzysztof Najman, Formal quality assessment of group structure mapping on the Kohonen's map 138
Kamila Migdał Najman, Krzysztof Najman, Graphical quality assessment of group structure mapping on the Kohonen's map 147
Beata Basiura, Anna Czapkiewicz, Validation of time series clustering 156
Michał Trzęsiok, Selected methods for outlier detection..... 166

Grażyna Dehnel, Tomasz Klimanek, Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics 176
Michał Bernard Pietrzak, Justyna Wilk, Economic distance in modeling spatial phenomena with the application of gravity model..... 185
Maciej Beręsewicz, An attempt to use different distance measures in the Generalized Petersen estimator 194
Marcin Szymkowiak, Tomasz Józefowski, Construction and practical using of SPREE estimators for two-dimensional contingency tables..... 201
Marcin Pełka, The ensemble conceptual clustering for symbolic data..... 209
Małgorzata Machowska-Szewczyk, Evaluation of clusters obtained by fuzzy classification methods for symbolic objects..... 219
Justyna Wilk, Problem of determining the number of clusters in taxonomic analysis of symbolic data 228
Andrzej Dudek, Clustering techniques for Google maps markers..... 236
Ewa Roszkowska, The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure 247
Marcin Szymkowiak, Marek Witkowski, The use of correspondence analysis in analysing the financial situation of cooperative banks..... 256
Bartłomiej Jefmański, The construction of fuzzy customer satisfaction indexes using R program..... 265
Karolina Bartos, Discovering patterns of consumer behaviour by market basket analysis of the transactional data..... 274
Joanna Trzęsiok, Cluster analysis of countries with respect to fertility rate and other demographic factors 284
Beata Bal-Domańska, An attempt to identify major regional clusters and their convergence 293
Beata Bieszk-Stolorz, Iwona Markowicz, The influence of benefit on the job finding process 302
Marta Dziechciarz-Duda, Klaudia Przybysz, Education and labor market needs. Classification of university graduates 312
Tomasz Klimanek, The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods..... 320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Selected methods for an analysis of longitudinal data..... 329
Artur Zaborski, The application of distance measures for ordinal data for aggregation individual preferences 337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market 345
Katarzyna Wawrzyniak, Classification of voivodeships according to the level of the realization of priorities of *the National Development Strategy 2007-2015* with using the values of centroid of the rows 355

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Artur Zaborski
Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE MIAR ODLEGŁOŚCI DLA DANYCH PORZĄDKOWYCH DO AGREGACJI PREFERENCJI INDYWIDUALNYCH

Streszczenie: W artykule dokonano klasyfikacji metod agregacji preferencji indywidualnych oraz przedstawiono metody wykorzystujące miary odległości. Wskazano miary, które mogą być stosowane do pomiaru odległości między relacjami preferencji różnych respondentów. Opisano miary wykorzystujące jedynie rozkłady preferencji dla wszystkich par obiektów, miary oparte na rangach oraz miarę GDM2, która w swojej konstrukcji wykorzystuje relacje dopuszczalne na skali porządkowej. W części empirycznej przedstawiono przykład, w którym agregację preferencji indywidualnych przeprowadzono z wykorzystaniem funkcji BruteAggreg programu R.

Słowa kluczowe: preferencje indywidualne, metody agregacji, miary odległości, środowisko R.

1. Wstęp

Agregacja indywidualnych ocen preferencji pewnego zbioru alternatyw umożliwia porównanie ich ze społecznego punktu widzenia, wyboru z tego zbioru alternatywy najlepszej lub podzbioru najlepszych alternatyw.

Do tej pory powstało wiele metod agregacji preferencji. Są to głównie metody wypracowane w ramach teorii wyboru społecznego (metody związane z zasadą zwykłej większości, metody związane z regułą Bordy, metoda Condorceta, metoda optymalnej predykcji i in.).

Celem pracy jest wskazanie możliwości wykorzystania do agregacji miar odległości między ocenami preferencji indywidualnych (np. odległość Spearmana, odległość τ – Kendalla) oraz miar stosowanych do pomiaru odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej.

W części empirycznej dokonano agregacji indywidualnych preferencji ekspertów zajmujących się sprzedażą detaliczną i naprawą sprzętu komputerowego względem wybranych marek monitorów LCD, za pomocą funkcji `BruteAggreg` programu R.

2. Preferencje indywidualne

W celu uporządkowania zbioru obiektów $X = \{x_1,\ldots,x_i,\ldots,x_m\}$ ze względu na preferencje stosuje się relacje preferencji mocnej ($x_i \succ x_j$), preferencji słabej ($x_i \succeq x_j$) oraz indyferencji ($x_i \approx x_j$). Jeżeli istnieje funkcja, która umożliwia pomiar obiektów na skali porządkowej, to wymienione relacje można przedstawić jako [por. Bąk 2004, s. 37]:

- $x_i \succ x_j \Leftrightarrow u(x_1) > u(x_2)$,
- $x_i \succeq x_j \Leftrightarrow u(x_1) \geq u(x_2)$,
- $x_i \approx x_j \Leftrightarrow u(x_1) = u(x_2)$,

gdzie funkcja *u* jest funkcją użyteczności, porządkującą analizowane obiekty zgodnie z preferencjami konsumenta. W badaniu preferencji nie są istotne wartości różnic między wartościami funkcji użyteczności poszczególnych konsumentów, przez co dozwolonymi przekształceniami matematycznymi dla obserwacji są ściśle monotoniczne funkcje rosnące, które nie zmieniają dopuszczalnych relacji, tj. równości, różności, większości i mniejszości.

Relacje preferencji indywidualnych powinny spełniać następujące warunki [zob. np. Varian 1997, s. 66; Bąk 2004, s. 36]:

- zwrotności – jeżeli dla każdego $x_i \in X$ zachodzi relacja $x_i \succeq x_i$, tzn. dwa identyczne obiekty nie są rozróżniane na skali preferencji danego konsumenta,
- spójności – dla każdej pary obiektów $x_i, x_j \in X$ musi być spełniony przynajmniej jeden z warunków: $x_i \succeq x_j \vee x_i \preceq x_j \vee x_i \approx x_j$,
- przechodniości – jeżeli dla każdej trójki obiektów $x_i, x_j, x_k \in X$ oceny konsumenta spełniają warunek racjonalności, tzn.: $x_i \succeq x_j \wedge x_j \succeq x_k \Rightarrow x_i \succeq x_k$.

Oceny formułowane przez konsumentów zazwyczaj spełniają te warunki. Warunek spójności może nie być spełniony w przypadku, gdy obiekty znacznie różnią się od siebie i trudno je umieścić na „wspólnej skali", zaś warunek przechodniości – gdy obiekty różnią się bardzo mało lub są na tyle złożone, że trudno je porównywać między sobą.

Informację o relacji preferencji *h*-tego respondenta otrzymuje się, prosząc go o uporządkowanie obiektów zbioru *X* od najbardziej do najmniej preferowanego. Można też poprosić respondenta o dokonanie, zgodnie z jego własnymi preferencjami, porównań wszystkich par obiektów. Ten drugi sposób jest jednak bardzo pracochłonny, zwłaszcza przy dużej liczbie obiektów. Ponadto w wyniku porównań parami, można otrzymać relację, która nie spełnia warunku przechodniości.

3. Klasyfikacja metod agregacji preferencji

Klasyfikacji metod agregacji preferencji można dokonać na podstawie dwóch kryteriów. Pierwsze z nich określa, jakie są wykorzystywane informacje o preferencjach indywidualnych. Według tego kryterium wyróżnia się dwa rodzaje metod:
- metody binarne – wykorzystujące jedynie rozkłady preferencji dla wszystkich par obiektów (np. otrzymanych na podstawie porównań parami),
- metody niebinarne – korzystające z pełniejszej informacji o relacjach preferencji (np. opierające się na uporządkowaniach preferencji).

Drugie kryterium klasyfikacji określa sposób, w jaki dokonuje się agregacji. Według tego kryterium możemy rozróżnić trzy grupy metod:
- miary tendencji centralnej – choć są najczęściej wykorzystywane, to taki sposób agregacji nie zawsze jest właściwy; mimo że preferencje są mierzone na skali porządkowej, to stosując te metody, często przyjmuje się założenie, że preferencje konsumentów mierzone są co najmniej na skali przedziałowej;
- metody wypracowane w ramach teorii wyboru społecznego – można tu wymienić metody związane z zasadą zwykłej większości (metoda Copelanda, metoda Tody), grupę metod związanych z regułą Bordy, metodę Condorceta, metodę optymalnej predykcji i in. [zob. Lissowski 2000];
- metody wykorzystujące miary odległości między indywidualnymi relacjami preferencji.

4. Agregacja preferencji z wykorzystaniem wybranych miar odległości

Agregacja preferencji indywidualnych z wykorzystaniem funkcji odległości polega na znalezieniu spośród permutacji uporządkowań należących do zbioru $\boldsymbol{Q}$, takiej relacji preferencji R^1, dla której suma odległości od wszystkich indywidualnych uporządkowań preferencji jest najmniejsza, tzn.:

$$\sum_{h=1}^{n} d(R_h, R^1) = \min_{R \in Q} \sum_{h=1}^{m} d(R_h, R), \tag{1}$$

gdzie: $d(R_h, R^1)$ – odległość między relacją preferencji *h*-tego respondenta (R_h) a R^1,
$\boldsymbol{Q}$ – zbiór wszystkich możliwych uporządkowań preferencji *m* obiektów.

Ponieważ mediana jest tą wartością, która minimalizuje sumę odległości wartości zmiennej od stałej, dlatego R^1 określa się medianą uporządkowań preferencji.

Drugą metodą wyznaczania zagregowanego uporządkowania preferencji jest wybór takiego, które minimalizuje sumę kwadratów odległości od indywidualnych uporządkowań, tzn.:

$$\sum_{h=1}^{n}\left[d(R_h,R^2)\right]^2=\min_{R\in Q}\sum_{h=1}^{m}d\left[(R_h,R)\right]^2. \tag{2}$$

Uporządkowanie R^2 nazywane jest średnią uporządkowań indywidualnych, ponieważ właśnie średnia minimalizuje sumę kwadratów odległości zmiennej od stałej.

Miary odległości między uporządkowaniami preferencji można podzielić na te, które wykorzystują binarne relacje preferencji (tzn. czy respondent przedkłada x_i nad x_j, czy x_j nad x_i, czy też jest wobec nich indyferentny) oraz miary oparte na rangach, w tym miary stosowane do pomiaru odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej.

Za najważniejszą binarną miarę odległości uznaje się miarę Kemeny'ego [Kemeny, Snell 1962]:

$$d(R_g,R_h)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\left|r_{ij}^{g}-r_{ij}^{h}\right|, \tag{3}$$

gdzie: $i,j=1,2,\ldots,m$ – numer obiektu,

$g,h=1,2,\ldots,n$ – numer respondenta,

$r_{ij}^{g(h)}=\begin{cases}1 \text{ gdy } x_i\succ x_j,\\ 0 \text{ gdy } x_i\prec x_j\vee x_i\approx x_j\end{cases}$ dla *g*-tego (*h*-tego) respondenta.

Odległość Kemeny'ego spełnia trzy postulaty metryki, a ze względu na swoją konstrukcję nazywana jest odległością „miejską".

Inną miarę odległości między uporządkowaniami preferencji wprowadził Bogart [Bogart 1973]:

$$d\left(R_g,R_h\right)=\frac{1}{\sqrt{2}}\left\|\mathbf{A}(R_g)-\mathbf{A}(R_h)\right\|, \tag{4}$$

gdzie: $\|\mathbf{A}\|$ – pierwiastek kwadratowy sumy kwadratów elementów macierzy **A**,

$\mathbf{A}(R_g)$ $\left(\mathbf{A}(R_h)\right)$ – macierz ocen preferencji *g*-tego (*h*-tego) respondenta o elementach $a_{ij}^{g(h)}=\begin{cases}1 \text{ gdy } x_i\succ x_j\\ 0 \text{ gdy } x_i\approx x_j\\ -1 \text{ gdy } x_i\prec x_j\end{cases}$.

Ze względu na postać i własności miarę Bogarta określa się jako miarę „euklidesową".

Spośród miar odległości typowych dla uporządkowań preferencji opartych na rangach można wymienić odległość Spearmana (*Spearman footrule distance*) oraz odległość τ – Kendalla [zob. Pihur, Datta, Datta 2009].

Odległość Spearmana przyjmuje postać:

$$d_S(R_g, R_h) = \sum_{i=1}^{m} \left| r^g(x_i) - r^h(x_i) \right|, \tag{5}$$

gdzie: $r^g(x_i)$ ($r^h(x_i)$) – ranga *i*-tego obiektu w profilu preferencji *g*-tego (*h*-tego) respondenta,

Odległość Spearmana może być znormalizowana tak, aby przyjmowała wartości z przedziału $[0;1]$. W tym celu wyrażenie (5) należy podzielić przez $m^2/2$.

Odległość τ – Kendalla [Kendall 1938] oparta jest na liczbie inwersji występujących w danej relacji preferencji w porównaniu z inną relacją preferencji. Odległość τ – Kendalla wyrażona jest wzorem:

$$d_K(R_g, R_h) = \sum_{i,j=1}^{m} K_{ij}\,, \tag{6}$$

gdzie:

$$K_{ij} = \begin{cases} 0 \text{ gdy } \left(r^g(x_i) < r^g(x_j) \wedge r^h(x_i) < r^h(x_j)\right) \vee \left(r^g(x_i) > r^g(x_j) \wedge r^h(x_i) > r^h(x_j)\right) \\ 1 \text{ gdy } \left(r^g(x_i) > r^g(x_j) \wedge r^h(x_i) < r^h(x_j)\right) \vee \left(r^g(x_i) < r^g(x_j) \wedge r^h(x_i) > r^h(x_j)\right) \end{cases}.$$

Podobnie jak dla odległości Spearmana odległość τ – Kendalla można znormalizować tak, aby jej wartości mieściły się w przedziale [0;1]. Normalizacji dokonuje się przez podzielenie wyrażenia (6) przez $m(m-1)/2$.

Do agregacji preferencji indywidualnych można również wykorzystać konstrukcje miar stosowanych do pomiaru odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej (zarówno bazujących na rangach, jak i miary GDM2, która wykorzystuje dopuszczalne relacje na skali porządkowej). Ponieważ dla różnych indywidualnych relacji preferencji poszczególnym obiektom mogą zostać przyporządkowane takie same oceny, dlatego możliwe jest stosowanie tylko tych miar, które dopuszczają rangi powiązane.

Jedną z takich miar jest odległość Podaniego [Podani 1999]. Odległość między dwoma relacjami preferencji wyrażonymi za pomocą rang przedstawia równanie:

$$d_p\left(R_g, R_h\right) = \sum_{i=1}^{m} \left(1 - \frac{\left| r^g(x_i) - r^h(x_i) \right| - \left(t_{gi} - 1\right)/2 - \left(t_{hi} - 1\right)/2}{R_i - \left(t_{i.\max} - 1\right)/2 - \left(t_{i.\min} - 1\right)/2} \right), \tag{7}$$

gdzie: t_{gi} (t_{hi}) – liczba respondentów, którzy przypisali taką samą rangę jak *g*-ty (*h*-ty) respondent *i*-temu obiektowi (łącznie z respondentem *g* (*h*)),

$t_{i.\max}$ ($t_{i.\min}$) – liczba respondentów, którzy przypisali maksymalną (minimalną) rangę *i*-temu obiektowi,

R_i – rozstęp wyznaczony na podstawie porangowanych wartości dla *i*-tego obiektu.

Miarą, która w swojej konstrukcji wykorzystuje relacje dopuszczalne na skali porządkowej, tj. równości, różności, większości i mniejszości, jest zaproponowana przez Walesiaka [Walesiak 1993, s. 44-45] miara GDM2. Miara GDM2 zastosowana do wyznaczania odległości między uporządkowaniami preferencji przyjmie postać:

$$d_w\left(R_g, R_h\right) = \frac{1}{2} - \frac{\sum_{i=1}^{m} a_{ghi} b_{ghi} + \sum_{i=1}^{m} \sum_{\substack{l=1 \\ l \neq g,h}}^{n} a_{gli} b_{hli}}{2\left[\sum_{i=1}^{m}\sum_{l=1}^{n} a_{gli}^2 \cdot \sum_{i=1}^{m}\sum_{l=1}^{n} b_{hli}^2\right]^{\frac{1}{2}}},$$

gdzie: $a_{gpi}\ \left(b_{hsi}\right) = \begin{cases} 1 \text{ gdy } x_{gi} \succ x_{pi}\ \left(x_{hi} \succ x_{si}\right) \\ 0 \text{ gdy } x_{gi} \approx x_{pi}\ \left(x_{hi} \approx x_{si}\right), \text{ dla } p = h,l;\ s = g,l, \\ -1 \text{ gdy } x_{gi} \prec x_{pi}\ \left(x_{hi} \prec x_{si}\right) \end{cases}$

$x_{gi}\ \left(x_{hi},\ x_{li}\right)$ – ocena preferencji i-tego obiektu przez g-tego (h-tego, l-tego) respondenta,

$g,h,l = 1,\ \ldots,\ n$ – numer respondenta,

$i = 1,\ \ldots,\ m$ – numer obiektu.

5. Agregacja preferencji indywidualnych w programie R

W programie R agregacja preferencji z wykorzystaniem miar odległości jest możliwa za pomocą funkcji `BruteAggreg` pakietu `RankAggreg`. Pomiar odległości między relacjami preferencji w funkcji `BruteAggreg` jest dokonywany z wykorzystaniem odległości Spearmana oraz odległości τ – Kendalla. Składnię funkcji oraz jej podstawowe argumenty prezentuje tab. 1.

Tabela 1. Opis funkcji `BruteAggreg` w programie R

`BruteAggreg(x, k, weights=NULL, distance=c("Spearman", "Kendall"), importance=rep(1,nrow(x)))`	
`x`	macierz uporządkowanych preferencji
`k`	liczba najważniejszych uporządkowań podlegających agregacji
`weights`	wagi uporządkowań preferencji podlegających agregacji
`distance`	wykorzystywana miara odległości
`importance`	wektor wag wskazujący ważność każdego uporządkowania preferencji

Źródło: opracowanie własne z wykorzystaniem dokumentacji programu R.

Przykład

Wybranym 28 ekspertom zajmującym się sprzedażą detaliczną, serwisowaniem i naprawą sprzętu komputerowego przedstawiono 8 marek monitorów LCD (Samsung, LG, Maxdata, Philips, Benq, NEC, Neovo, Hyundai) z prośbą o uszeregowanie swoich preferencji poprzez przyporządkowanie poszczególnym markom rang od 1 do 8, przy czym liczba 1 oznaczała markę najbardziej preferowaną. Następnie, wykorzystując skrypt 1, dokonano agregacji ocen preferencji za pomocą funkcji `BruteAggreg`:

Skrypt 1

```
library(RankAggreg)
x<-read.csv2("monitory_pref.csv", header=TRUE)
liczbaObiektow<-ncol(x)
x<-as.matrix(x)
m1<-BruteAggreg(x, liczbaObiektow, distance="Kendall")
m2<-BruteAggreg(x, liczbaObiektow, distance="Spearman")
print(m1, quote=FALSE)
print(m2, quote=FALSE)
plot(m1)
plot(m2)
```

W wyniku zastosowania skryptu 1 otrzymano zagregowane uporządkowanie ocen preferencji oddzielnie dla odległości Spearmana oraz odległości τ – Kendalla:

```
Algorithm: BruteForce
 Distance: Kendall
 Score: 6.714286
The optimal list is:
 Samsung Philips LG Benq NEC Hyundai Maxdata Neovo

 Distance: Spearman
 Score: 11.42857
The optimal list is:
 Samsung Philips LG Benq NEC Hyundai Neovo Maxdata
```

6. Podsumowanie

W artykule przedstawiono metodę agregacji preferencji indywidualnych z wykorzystaniem miar odległości. Wskazano miary, które mogą być stosowane do pomiaru odległości między relacjami preferencji różnych respondentów. Przedstawiono miary wykorzystujące jedynie rozkłady preferencji dla wszystkich par obiektów (np. otrzymanych na podstawie porównań parami), miary oparte na rangach oraz miarę GDM2, która w swojej konstrukcji wykorzystuje relacje dopuszczalne na skali porządkowej.

W części empirycznej przedstawiono przykład, w którym agregację preferencji indywidualnych przeprowadzono z wykorzystaniem funkcji `BruteAggreg` programu R. Pomiaru odległości między relacjami preferencji dokonano za pomocą odległości Spearmana oraz odległości τ – Kendalla, ponieważ jedynie te dwie miary są stosowane w funkcji `BruteAggreg`. W dalszych pracach zostaną podjęte próby rozszerzenia oprogramowania o inne miary odległości, dzięki czemu możliwa będzie również agregacja indyferentnych relacji preferencji.

Literatura

Bąk A. (2004), *Dekompozycyjne metody pomiaru preferencji w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej, Wrocław.

Bogart K.J. (1973), *Preference Structures I: Distances Between Trasitive Asymetric Relations*, „Journal of Mathematical Sociology", no. 3, s. 49-67.

Kemeny J.G., Snell L. (1962), *Mathematical Models in the Social Sciences*, Ginn, Boston, s. 9-23.

Kendall M.G. (1938), *A new measure of rank correlation*, „Biometrika", no. 30.

Lissowski G. (2000), *Metody agregacji indywidualnych preferencji*, „Studia Socjologiczne", nr 1, 2.

Pihur V., Datta S., Datta S. (2009), *RankAggreg, an R package for weighted rank aggregation*, BMC Bioinformatics, http://www.biomedcentral.com/1471-2105/10/62.

Podani J. (1999), *Extending gowers general coefficient of similarity to ordinal characters*, „Taxon", no 48.

Varian H.R. (1997), *Mikroekonometria*, PWN, Warszawa.

Walesiak M. (1993), *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654, Monografie i Opracowania nr 101, Wydawnictwo Akademii Ekonomicznej, Wrocław.

THE APPLICATION OF DISTANCE MEASURES FOR ORDINAL DATA FOR AGGREGATION INDIVIDUAL PREFERENCES

Summary: The article presents the classification of individual preferences aggregation methods and shows the methods which use the distance measures. Measures which can be applied to measuring the distance between different respondents preference relationship were discussed. The article describes measures which use preference distributions for all pairs of objects (e.g. obtained from pairwise comparisons), measures based on ranks and distance measure using permissible transformations to ordinal scale (GDM2 distance). In the empirical part the example of individual preference aggregation was carried out by BruteAggreg function of R program.

Keywords: individual preferences, aggregation methods, distance measures, R software.