EKONOMETRIA ECONOMETRICS 4(38) • 2012

ISSN 1507-3866

Artur Zaborski

Uniwersytet Ekonomiczny we Wrocławiu

AGREGACJA PREFERENCJI INDYWIDUALNYCH Z WYKORZYSTANIEM MIAR ODLEGŁOŚCI I PROGRAMU R

Streszczenie: Celem artykułu jest zaprezentowanie metod agregacji indywidualnych ocen preferencji za pomocą dwóch miar odległości. Są to odległość Spearmana (*Spearman footrule distance*) i odległość τ – Kendalla. Miary te są wykorzystywane do agregacji rang w środowisku R. Na zakończenie zaprezentowano przykład agregacji preferencji z wykorzystaniem funkcji `BruteAggreg` pakietu `RankAggreg`.

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

1. Wstęp

Agregacja indywidualnych ocen preferencji jest jednym z podstawowych problemów teorii wyboru społecznego. Najbardziej znaną metodą agregacji jest metoda Bordy, w której łącznego uporządkowania relacji preferencji dokonuje się na podstawie pozycji, jakie zajmują obiekty w indywidualnych uporządkowaniach preferencji poszczególnych osób.

Pomocnym narzędziem agregacji preferencji mogą być miary odległości między uporządkowaniami preferencji obiektów dokonanymi przez poszczególnych respondentów. Należy jednak pamiętać, że w tym celu dopuszczalne jest stosowanie wyłącznie takich miar odległości, które są właściwe dla zmiennych mierzonych na skali porządkowej.

W artykule zaprezentowano ideę agregacji preferencji z wykorzystaniem miar odległości bazujących na rangach. Szczególną uwagę zwrócono na miarę odległości Spearmana oraz τ – Kendalla, ponieważ te dwie miary są wykorzystywane w programie R do agregacji rang.

W części empirycznej przedstawiono przykład, w którym agregacji preferencji indywidualnych dokonano za pomocą funkcji `BruteAggreg` programu R.

2. Preferencje indywidualne i zasada zwykłej większości

Niech $X = \{x_1, \ldots, x_i, \ldots, x_m\}$ oznacza zbiór *m* obiektów poddawanych ocenie preferencji, zaś $N = \{1, \ldots h, \ldots, n\}$ jest zbiorem respondentów (konsumentów) dokonujących ocen preferencji. Preferencje indywidualne respondenta *h* $(h = 1, 2, \ldots, n)$ reprezentowane są przez binarną relację $x_i P_h x_j$, co oznacza, że obiekt x_i jest przynajmniej tak samo preferowany przez osobę *h* jak obiekt x_j.

W celu uporządkowania zbioru obiektów *X* ze względu na preferencje stosuje się następujące relacje:

– preferencji mocnej: $x_i \succ x_j$,

– preferencji słabej: $x_i \succeq x_j$,

– indyferencji: $x_i \approx x_j$.

Jeżeli istnieje funkcja, która umożliwia pomiar obiektów na skali porządkowej, to wymienione relacje można przedstawić w następujący sposób (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: $x_i \succeq x_j \vee x_i \preceq x_j \vee x_i \approx x_j$, tzn. konsument musi być w stanie rozstrzygnąć, czy przedkłada x_i nad x_j, czy x_j nad x_i, czy też jest wobec nich indyferentny,

– 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", warunek przechodniości zaś – gdy obiekty różnią się bardzo mało lub są na tyle złożone, że trudno jest je porównywać między sobą.

Informację o relacji preferencji *h*-tego respondenta (R_h) otrzymuje się, prosząc go o uporządkowanie obiektów zbioru *X* od najbardziej do najmniej preferowanego. Na przykład relacja $R_h : x_1, x_2 - x_3$ oznacza, że $x_1 \succ x_2 \approx x_3$. 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 spójności lub przechodniości.

Relacja preferencji *h*-tego respondenta może być również prezentowana w postaci macierzy o wymiarach $m \times m$. Wiersze i kolumny tej macierzy odpowiadają poszczególnym obiektom $x_i \in X$, zaś elementami tej macierzy są liczby $r_{ij}^h \in \{0,1\}$, przy czym:

$$r_{ij}^h = \begin{cases} 1 \text{ gdy } x_i \succ x_j, \\ 0 \text{ gdy } x_i \prec x_j \vee x_i \approx x_j \end{cases}. \tag{1}$$

Na przykład relacja $R_h : x_1, x_2 - x_3$ w zapisie macierzowym ma postać:

$$R_h = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Dla *n* respondentów profilem preferencji indywidualnych R^n będziemy nazywali funkcję $\phi : N \to \boldsymbol{Q}$, która każdej osobie ze zbioru *N* przypisuje pewną relację (uporządkowanie) preferencji obiektów *X* należącą do $\boldsymbol{Q}$ ($\boldsymbol{Q}$ – zbiór wszystkich możliwych relacji preferencji elementów zbioru *X*). Jest to ciąg *n* indywidualnych relacji preferencji:

$R^n = \{R_1, \ldots, R_h, \ldots, R_n\}$.

Przykład profilu preferencji indywidualnych dla $m = 4$ i $n = 4$ prezentuje tab. 1.

Wiele metod agregacji preferencji opiera się na określeniu rozkładów preferencji dla wszystkich możliwych par (x_i, x_j), polegającym na podaniu liczby osób, które:

– przedkładają x_i nad x_j, tj. $N(x_i \succ x_j)$,

– przedkładają x_j nad x_i, tj. $N(x_j \succ x_i)$,

– są indyferentne względem x_i oraz x_j, tj. $N(x_i \approx x_j)$.

Tabela 1. Przykładowy profil preferencji ($m = 4$ i $n = 3$)

Numer respondenta	Relacja preferencji indywidualnej
1	x_1, x_2, x_3, x_4
2	x_3, x_1, x_2, x_4
3	x_4, x_3, x_2, x_1
4	x_2, x_3, x_1, x_4

Źródło: opracowanie własne.

Prezentacja rozkładów preferencji indywidualnych dla wszystkich par (x_i, x_j) jest możliwa za pomocą macierzy:

$$R_\Sigma = \sum_{h=1}^{n} R_h = \left[r_{ij}^{\Sigma} \right]. \tag{2}$$

Elementy tej macierzy (r_{ij}^{Σ}) określają liczbę osób, dla których $x_i \succ x_j$.

Chociaż elementy macierzy R_Σ informują o liczbie osób, dla których między x_i oraz x_j zachodzi wyłącznie relacja mocnej preferencji, to na jej podstawie można obliczyć liczbę osób indyferentnych względem x_i oraz x_j według wzoru:

$$N(x_i \approx x_j) = n - N(x_i \succ x_j) - N(x_i \prec x_j). \tag{3}$$

Dla profilu preferencji indywidualnych przedstawionych w tab. 1 R_Σ przyjmuje postać:

$$R_\Sigma = \begin{bmatrix} 0 & 2 & 1 & 3 \\ 2 & 0 & 2 & 3 \\ 3 & 2 & 0 & 3 \\ 1 & 1 & 1 & 0 \end{bmatrix}.$$

Jedną z najstarszych metod agregacji indywidualnych preferencji jest zasada zwykłej większości. Jej podstawą są rozkłady preferencji dla wszystkich par badanych obiektów. Opiera się ona na założeniu, że:

$$x_i \succ x_j \Leftrightarrow N(x_i \succ x_j) > N(x_j \succ x_i),$$
$$x_i \approx x_j \Leftrightarrow N(x_i \succ x_j) = N(x_j \succ x_i),$$
$$x_i \prec x_j \Leftrightarrow N(x_i \succ x_j) < N(x_j \succ x_i).$$

Istnieje wiele metod związanych z zasadą zwykłej większości (zob. np. [Lissowski 2000]). Wśród nich można wymienić m.in. metodę Copelanda oraz metodę Bordy. Pierwsza z nich wykorzystuje relacje preferencji dla wszystkich par obiektów, druga zaś – pełne profile preferencji poszczególnych respondentów.

W metodzie Copelanda [1951] wprowadza się wskaźnik liczbowy, który każdemu obiektowi x_i przyporządkowuje, zgodnie z zasadą zwykłej większości, różnicę między liczbą x_i, dla których $x_i \succ x_j$, a liczbą x_i, dla których $x_i \prec x_j$, tj.:

$$c(x_i) = \left|\left\{x_i : N(x_i \succ x_j) > N(x_i \prec x_j)\right\}\right| - \left|\left\{x_i : N(x_i \succ x_j) < N(x_i \prec x_j)\right\}\right|. \tag{4}$$

Wartości tego wskaźnika stanowią podstawę do uporządkowania obiektów.

W metodzie Bordy obiektom przypisywane są pozycje (rangi), jakie zajmują one w indywidualnych uporządkowaniach preferencji. Łączna liczba punktów otrzymanych przez obiekty ze względu na ich pozycje we wszystkich indywidualnych relacjach preferencji decyduje o łącznym uporządkowaniu obiektów. Odmiany metod związanych z regułą Bordy prezentuje Lissowski [2000].

3. Wykorzystanie miar odległości do agregacji preferencji indywidualnych

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

$$\sum_{h=1}^{n} d(R_h, R^*) = \min_{R \in Q} \sum_{h=1}^{m} d(R_h, R), \tag{5}$$

gdzie: $d(R_h, R^*)$ – odległość między R_h a R^*,

$\boldsymbol{Q}$ – zbiór wszystkich możliwych uporządkowań preferencji *m* obiektów.

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

Do pomiaru odległości między relacjami preferencji dla poszczególnych respondentów mają zastosowanie miary odległości bazujące na rangach. Przed zastosowaniem właściwych miar relacje preferencji poszczególnych respondentów zostają porangowane. Na przykład dla relacji preferencji $R_h : x_2, x_3, x_1, x_4$ wartości rang poszczególnych obiektów wynoszą: $r^h(x_1) = 3$, $r^h(x_2) = 1$, $r^h(x_3) = 2$ i $r^h(x_4) = 4$ ($r^h(x_i)$ – ranga obiektu x_i w indywidualnej relacji preferencji dla *h*-tego respondenta) .

Spośród wielu miar opartych na rangach (zob. np. [Walesiak 2012]) można wymienić odległość Spearmana (*Spearman footrule distance*) oraz odległość τ – Kendalla[1] (zob. [Pihur, Datta, Datta 2009]).

Odległość Spearmana przyjmuje postać:

$$d_S(R_g, R_h) = \sum_{i=1}^{m} \left| r^g(x_i) - r^h(x_i) \right|, \tag{6}$$

gdzie: $r^g(x_i)$ $(r^h(x_i))$ – ranga i-tego obiektu w profilu preferencji g-tego (h-tego) respondenta,
$i = 1, 2, \ldots, m$ – numer obiektu,
$g, h = 1, 2, \ldots, n$ – numer respondenta.

Odległość Spearmana może być znormalizowana tak, aby przyjmowała wartości z przedziału $[0;1]$. W tym celu wyrażenie (6) należy podzielić przez $m^2/2$.

Przykład 1

Respondenci przedstawili swoje preferencje odnośnie do korzystania ze środków transportu publicznego (zob. tab. 2). Wektory relacji preferencji dla poszczególnych respondentów przedstawiały się następująco: (samolot $\succ$ pociąg $\succ$ samochód $\succ$ autobus) oraz (samochód $\succ$ pociąg $\succ$ autobus $\succ$ samolot).

Tabela 2. Relacje preferencji dla dwóch respondentów

Respondenci	Rangi			
	1	2	3	4
R_1	samolot	pociąg	samochód	autobus
R_2	samochód	pociąg	autobus	samolot

Źródło: dane umowne.

Znormalizowana odległość Spearmana między przedstawionymi relacjami preferencji wynosi:

$$d_S(R_1, R_2) = \frac{|1-4| + |2-2| + |3-1| + |4-3|}{8} = 0,75.$$

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

$$d_K(R_g, R_h) = \sum_{i,j=1}^{m} K_{ij}, \tag{7}$$

[1] Wybór miar odległości został podyktowany ich wykorzystaniem w programie R.

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 (7) przez $m(m-1)/2$.

W celu obliczenia wartości odległości τ – Kendalla między relacjami preferencji dla przedstawionego wcześniej przykładu należy dla każdej pary obiektów obliczyć wartość K_{ij} (zob. tab. 3).

Tabela 3. Inwersje w porównywanych relacjach preferencji

Pary obiektów	R_1	R_2	K_{ij}
(samolot, pociąg)	$1 < 2$	$4 > 2$	1
(samolot, samochód)	$1 < 3$	$4 > 1$	1
(samolot, autobus)	$1 < 4$	$4 > 3$	1
(pociąg, samochód)	$2 < 3$	$2 > 1$	1
(pociąg, autobus)	$2 < 4$	$2 < 3$	0
(samochód, autobus)	$3 < 4$	$1 < 3$	0

Źródło: opracowanie własne.

Znormalizowana odległość τ – Kendalla dla przedstawionych relacji preferencji wynosi:

$$d(R_1, R_2) = \frac{4}{4(4-1)/2} = 0,67.$$

4. Agregacja preferencji w programie R

W programie R agregacja preferencji z wykorzystaniem funkcji bazujących na rangach jest możliwa za pomocą funkcji `BruteAggreg` pakietu `RankAggreg`. Składnię funkcji oraz jej podstawowe argumenty prezentuje tab. 4.

Tabela 4. 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 2

Na podstawie zbudowanego drzewa klasyfikacyjnego z wykorzystaniem algorytmu CHAID (wartości podstawowych parametrów algorytmu ustalono następująco: *Maximum Tree Depth* = 4, *Parent Node* = 45, *Child Node* = 15) wyodrębniono na podstawie poziomu zaawansowania w nauce języka obcego sześć względnie jednorodnych klas słuchaczy jednej ze szkół językowych (szczegółowe wyniki klasyfikacji przedstawiono w pracy [Kurzydłowski, Zaborski 2004]). Ze zbioru zmiennych objaśniających w konstrukcji drzewa zostały uwzględnione zmienne charakteryzujące wykształcenie respondenta, jego miejsce nauki oraz poznawany język obcy. Pozostałe zmienne objaśniające, które brano pod uwagę w analizie, okazały się statystycznie nieistotne ze względu na prowadzony cel badania, tzn. nie wystąpiła statystycznie istotna zależność między określoną zmienną objaśniającą a zmienną objaśnianą.

Respondentom przedstawiono 7 następujących czynników: poziom kształcenia (A), kwalifikacje kadry (B), stosowane metody nauczania (C), atmosferę w szkole (D), cenę (E), lokalizację szkoły (F) i opinie znajomych (G), oraz poproszono o uporządkowanie, zgodnie z własnymi preferencjami, ważności tych czynników przez przyporządkowanie im kolejnych liczb naturalnych od 1 do 7. Liczba 1 oznaczała czynnik najbardziej istotny, liczba 7 zaś – czynnik mający dla respondentów najmniejsze znaczenie.

Oddzielnie dla każdej klasy słuchaczy dokonano agregacji ocen preferencji z wykorzystaniem funkcji `BruteAggreg` z następującą składnią poleceń:

```
library(RankAggreg)
x<-read.csv2("dane1.csv", header=FALSE)
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).
```

Wyniki analizy zestawiono w tab. 5.

Tabela 5. Zagregowane uporządkowanie czynników z wykorzystaniem odległości Spearmana i τ – Kendalla

Klasa	Odległość Spearmana		Odległość τ – Kendalla	
	Uporządkowanie	$\sum d_S(R^*, R_h)$	Uporządkowanie	$\sum d_K(R^*, R_h)$
I	A C B D E F G	8,411765	A C B D E F G	5,000000
II	C A B D E F G	7,285714	C A B D E F G	4,547619
III	C A B D E F G	8,875000	C A B D E F G	5,437500
IV	A C B D E F G	8,222222	A C B D F E G	4,888889
V	C A B D E F G	7,684211	C A B D E F G	4,491228
VI	C A B D E F G	7,657143	C A B D E F G	4,742857

Źródło: opracowanie własne z wykorzystaniem funkcji `BruteAggreg`.

Uzyskane wyniki agregacji wskazują, że bez względu na przynależność do określonej klasy uporządkowanie preferencji respondentów względem przedstawionych czynników jest prawie takie samo. Wyjątek stanowią słuchacze zaklasyfikowani do klas I i IV (osoby z wykształceniem podstawowym lub średnim oraz uczniowie szkół podstawowych), dla których poziom kształcenia (A) jest ważniejszy niż atmosfera panująca w szkole (C). Ponadto agregacja z wykorzystaniem odległości τ – Kendalla wskazała, że słuchacze klasy IV (uczniowie szkół podstawowych) przykładają nieco większą wagę do lokalizacji szkoły (F) niż pozostali.

5. Podsumowanie

W artykule zaprezentowano metodę agregacji preferencji indywidualnych z wykorzystaniem wybranych miar odległości. Skoncentrowano się na następujących miarach: odległości Spearmana, która jest sumą wartości bezwzględnej różnicy rang poszczególnych obiektów w dwóch uporządkowaniach preferencji, oraz odległości τ – Kendalla opartej na liczbie inwersji występujących w porównywanych relacjach preferencji. Miary te, chociaż nie są typowymi miarami dla zmiennych mierzonych na skali porządkowej (zob. [Walesiak 2012]), są stosowane do agregacji rang w programie R. Jednak zaleta programu R, jaką jest otwarty kod źródłowy ułatwiający rozwój procedur i dostosowanie ich do potrzeb użytkownika, umożliwia modyfikację funkcji agregacji np. o typową dla danych porządkowych miarę GDM2, której dodatkowym atutem jest możliwość zastosowania dla preferencji indyferentnych.

Literatura

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

Copeland A.H., *A reasonable Social Welfare Function*, University of Michigan, 1951.

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

Kurzydłowski A., Zaborski A., *Zastosowanie wybranych metod wielowymiarowej analizy statystycznej w badaniu preferencji*, [w:] Taksonomia 11, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1022, AE, Wrocław 2004.

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

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

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

Walesiak M., *Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategie postępowania*, [w:] Taksonomia 19, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 242, Wrocław 2012.

INDIVIDUAL PREFERENCES AGGREGATION BY USING DISTANCE MEASURES AND R PROGRAM

Summary: The aim of this paper is to present the methods of aggregating individual preferences scores by using two distance measures. They are Spearman footrule distance and Kendall's tau distance. These measures are used to rank aggregation in the R software. Finally, an example of preference aggregation is presented in which the `BruteAggreg` function of `RankAggreg` package is used.

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