A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 3(302), 2014

Artur Zaborski[*], *Marcin Pełka*[**]

DISTANCE MEASURES IN AGGREGATING PREFERENCE DATA

Abstract. The aim of this paper is to present aggregation methods of individual preferences scores by means of distance measures. Three groups of distance measures are discussed: measures which use preference distributions for all pairs of objects (e.g. Kemeny's measure, Bogart's measure), distance measures based on ranking data (e.g. Spearman distance, Podani distance) and distance measures using permissible transformations to ordinal scale (GDM2 distance). Adequate distance formulas are presented and the aggregation of individual preference by using separate distance measures was carried out with the use of the R program.

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

I. INTRODUCTION

The problem of preferences rank aggregation is the problem of computing a "consensus" ranking, given individual ranking preferences of several judges. Many aggregation methods have been proposed in the literature. They are mainly the methods developed within the theory of social choice. The article presents preference aggregation by using distance measures for ranking data. Due to the fact that the majority of these measures are not typical for ordinal data (see: Walesiak (2011)) it was checked if allowed mathematical transformations for ordinal data influence aggregation results.

II. INDIVIDUAL PREFERENCES

Let denote $X = \{x_1, \ldots, x_i, \ldots, x_m\}$ is the set of m objects, and $N = \{1, \ldots h, \ldots, n\}$ is the set of respondents (consumers) evaluating preferences. Personal preferences of the h-th respondent $(h = 1,2,\ldots,n)$ are represented by a binary relation $x_i P_h x_j$, which means that the object x_i is at least as preferred by the person h as the object x_j.

[*] Ph.D., Department of Econometrics and Computer Science, Wrocław Univeristy of Economics.
[**] Ph.D., Department of Econometrics and Computer Science, Wrocław Univeristy of Economics.

In order to organize the set of objects due to preferences we can apply a strong relation of preferences ($x_i \succ x_j$), a weak relation ($x_i \succeq x_j$) and an indifference ($x_i \approx x_j$). If there is a function that allows us to measure objects on an ordinal scale, than the mentioned relations can be presented as follows (see: Bąk (2004)):

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

where *u* is the function of utility which orders objects according to the consumers' preferences. In the preferences study the differences between the values of the utility function of each consumer are not significant, so allowed mathematical transformations for the observation are only strictly monotone increasing functions, which do not change the permissible relations for ordinal scale, i.e. equality, inequality, majority and minority.

III. THE CLASSIFICATION OF PREFERENCE AGGREGATION METHODS

The classification of preference aggregation methods can be made on the basis of two criteria. The first one determines which type of information about the individual preferences is used. According to this criterion there are two types of methods:

– binary methods – using only preferences decompositions for all pairs of objects (e.g. obtained by pairwise comparisons),

– non-binary methods – using fuller information about the preferences relationship (e.g. based on preferences rankings).

The second criterion of classification determines how the aggregation is done. According to this criterion, we can distinguish three groups of methods:

– the central tendency measures – despite being most common, they are not always appropriate. Preferences are measured on an ordinal scale, but when we use these methods it is often assumed that consumer preferences are measured at least on an interval scale;

– the methods developed within the theory of social choice – it is possible to list the methods associated with the majority rule (the Copeland's method , the Toda's method), a group of methods associated with the Borda's rule, Condorcet method, the method of optimal prediction et al. (see: Lissowski (2000));

– methods based on measurement of distances between individual preference relations.

IV. PREFERENCE AGGREGATION BY USING THE DISTANCE MEASURES

The idea of individual preferences aggregation with the use of the distance measures is to find such a relation of preference from all permutations of the orderings, for which the sum of distances from all individual preference orderings is the smallest, i.e.:

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

where: $d(R_h, R^1)$ – the distance between the preference relation of the h-th respondent (R_h) and R^1; Q – the set of all possible preference orderings of m objects.

Since the median is the value which minimizes the sum of the distances from the variables, therefore R^1 defines the median of preference orderings.

The second method of aggregation preference orderings is choosing such a relation that minimizes the sum of squared distances from individual orderings, i.e.:

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

R^2 is called the mean of individual orderings, because the mean minimizes the sum of square distances from the variables.

V. CHOSEN DISTANCE MEASURES FOR PREFERENCE AGGREGATION

Distance measures that can be applied for ordered preferences can be divided into two types – distance measures that use only binary relations between preferences (that show if respondent has chosen x_i over x_j, x_j over x_i or his choice is indifferent) and distance measures that use ranks as the input.

Kemeny's distance measure is the most important binary distance measure for preference data (Kemeny *et al*. (1962)):

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

where: $i, j = 1,2,\ldots,m$ – object number; $g, h = 1,2,\ldots,n$ – respondent number;

$$r_{ij}^{g(h)} = \begin{cases} 1 \text{ when } x_i \succ x_j, \\ 0 \text{ when } x_i \prec x_j \vee x_i \approx x_j \end{cases} \text{ for } g\text{-}th \text{ } (h\text{-th}) \text{ respondent.}$$

Kemeny's distance is a metric and due to its form is sometimes called „city block distance".

Bogart has proposed another important distance measure for ordered preferences that is sometimes called "Euclidean" (Bogart (1973)):

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

where: $\|\mathbf{A}\|$ – square root of the sum of the squared elements from the **A** matrix; $\mathbf{A}(R_g)\left(\mathbf{A}(R_h)\right)$ – matrix of preference evaluation for *g*-th (*h*-th) respondent with elements:

$$a_{ij}^{g(h)} = \begin{cases} 1 \text{ wheny } x_i \succ x_j \\ 0 \text{ when } x_i \approx x_j \\ -1 \text{ when } x_i \prec x_j \end{cases}.$$

There are many different distance measures for ordered rank data – most important are: Spearman's distance (also known as *Spearman footrule distance*) and τ – Kendall distance (see: Pihur *et al*. (2009)).

Spearman footrule distance is expressed by the following equation:

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

where: $r^g(x_i)$ $(r^h(x_i))$ – rank of *i*-th object for *g*-th (*h*-th) respondent.

Spearman's footrule distance can take its values from the interval $[0; 1]$ by normalization. To obtain such values equation (5) has to be divided by $m^2/2$.

τ – Kendall distance (Kendall (1938)) is based on the number of inversions occurring in the analysed preference relation in comparison with other preference relation. The τ – Kendall distance is expressed as follows:

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

where:

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

.

Like the *Spearman footrule distance*, τ – Kendall distance can take its value values from the interval $[0;1]$ after normalization of equation (6) by dividing it by $m(m-1)/2$.

To aggregate individual preference data also another distance measures can be applied that are usually used to calculate distances between objects described by ordinal variables. As different individuals (respondents) can carry out the same evaluations of some analyzed objects, only distance measure which allows to analyze tied preferences can be applied.

One of such distance measures is the Podani distance (Podani (1999)). The distance between two relations of preferences is expressed as follows:

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

where: t_{gi} (t_{hi}) – number of respondents that have assigned the same rank as the *g*-th (*h*-th) respondent for *i*-th object (including respondent *g* (*h*)); $t_{i.\max}$ ($t_{i.\min}$) – number of respondents that have assigned maximum (minimum) rank for *i*-th object; R_i – spread for ranked values for *i*-th object.

GDM2 (*General Distance Measure*) proposed by Walesiak (1993) is a distance measure that takes into account available relations for ordinal variables. GDM2 distance measure for ordered preference data is defined as follows:

$$d_w\left(R_g, R_h\right) = \frac{1}{2} - \frac{\sum_{i=1}^{m} a_{ghi} b_{ghi} + \sum_{i=1}^{m} \sum_{\substack{l=1 \\ l \neq g,h}}^{n} a_{gli} b_{hli}}{2\left[\sum_{i=1}^{m} \sum_{l=1}^{n} a_{gli}^2 \cdot \sum_{i=1}^{m} \sum_{l=1}^{n} b_{hli}^2\right]^{\frac{1}{2}}} \,, \tag{8}$$

where:

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

$x_{gi}\left(x_{hi}, x_{li}\right)$ – preference evaluation for *i*-th objects and *g*-th (*h*-th, *l*-th) respondent,

$g, h, l = 1, \ldots, n$ – respondent number,

$i = 1, \ldots, m$ – object number.

VI. EXAMPLE

Rank aggregation was made for two data sets: LCD brands data and sports data. LCD brands data contains eight different LCD brands – Samsung (Sa), LG (LG), Maxdata (Ma), Philips (Phi), BenQ (Ben), NEC (NE), Neovo (Neo), Hyundai (Hyu). 28 PC experts and dealers were asked to rank 8 LCD display brands according to their preferences on a 8-point scale: 1 – the highest preferred brand, 8 – the least preferred brand.

In sports data 130 students at the University of Illinois were asked to rank seven sports according to their preference in participating: Baseball (Bas), Football (Foo), Basketball (Bab), Tennis (Ten), Cycling (Cyc), Swimming (Swi), Jogging (Jog) (see: Marden (1995)).

Both data sets were applied in the experimental evaluation of four different distances for ordinal data in order to check the stability of the distances in rank data aggregation when dealing raw data, data after transformations ($y = x^2$ and $y = \sqrt{x}$). The results of evaluations are shown in Table 1.

When comparing two different data sets and four distances (see Table 1) one can notice that in the case of LCD brands data almost all distances applied reached almost the same final ranking results (there are some slight changes in the case of τ -Kendall distance) regardless of the transformation applied. In the case of sports data the changes between the τ -Kendall distance and other distance measures are more obvious than in the case of LCD brand data sets.

Table 1. Comparison of results for two different data sets and four different distance types

Distance	Transformation	LCD brands								Sports						
		Sa	LG	Ma	Phi	Ben	NE	Neo	Hyu	Bas	Foo	Bab	Ten	Cyc	Swi	Jog
d_1	$y = x$	1	3	7	2	4	5	8	6	2	1	3	4	5	7	6
	$y = x^2$	1	3	7	2	4	5	8	6	2	1	3	4	5	7	6
	$y = \sqrt{x}$	1	3	7	2	4	5	8	6	2	1	3	4	5	7	6
d_2	$y = x$	1	3	8	2	4	5	7	6	1	6	4	5	2	3	7
	$y = x^2$	1	3	8	2	4	5	7	6	1	6	4	5	2	3	7
	$y = \sqrt{x}$	1	3	8	2	4	5	7	6	1	6	4	5	2	3	7
d_3	$y = x$	1	3	8	2	4	5	7	6	1	6	4	5	2	3	7
	$y = x^2$	1	3	8	2	4	5	7	6	1	6	4	5	2	3	7
	$y = \sqrt{x}$	1	3	8	2	4	5	7	6	1	6	4	5	2	3	7
d_4	$y = x$	1	3	8	2	5	4	7	6	1	6	4	5	3	2	7
	$y = x^2$	1	3	8	2	5	4	7	6	1	6	4	5	3	2	7
	$y = \sqrt{x}$	1	3	8	2	5	4	7	6	1	6	4	5	3	2	7

where: d_1 – τ-Kendall distance; d_2 – *Spearman footrule distance*; d_3 – Podani distance; d_4 – GDM2 distance (GDM distance for ordinal variables).

Source: author's eleboration with application of R software.

VII. CONCLUSIONS

Combining the ranked preferences of many experts is an old and deep problem that has gained renewed importance in many applications. The use of distance measures for preferences aggregation is an alternative to methods developed within the theory of social choice. Although there are no typical distance measures for ordinary data (expect for GDM2) it was shown that for distance measures for rank data strictly monotone increasing functions permissible for ordinal scale do not influence aggregation results.

REFERENCES

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

Bogart K.J. (1973), Preference Structures I: Distances Between Transitive Asymmetric Relations, *Journal of Mathematical Sociology*, 3, 49–67.

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

Kendall M.G. (1938), A new measure of rank correlation, *Biometrika*, 30, 81–89.
Lissowski G. (2000), Metody agregacji indywidualnych preferencji, *Studia Socjologiczne*, 1–2, 79–103.
Marden J.I. (1995), *Analysis and modeling rank data*, Chapman and Hall, London.
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*, 48, 331–340.
Walesiak M. (1993), *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654. Seria: Monografie i opracowania nr 101, Wrocław.
Walesiak M. (1911), *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław.

Artur Zaborski, *Marcin Pełka*

MIARY ODLEGŁOŚCI W AGREGACJI DANYCH PREFERENCJI

Celem artykułu jest zaprezentowanie metod agregacji indywidualnych ocen preferencji za pomocą wybranych miar odległości. Omówiono trzy grupy miar odległości: miary wykorzystujące rozkłady preferencji dla wszystkich par obiektów (np. miara Kemeny'ego, miara Bogarta), miary odległości bazujące na rangach (np. odległość Spearman'a, odległość Podaniego) oraz miary odległości wykorzystujące dopuszczalne relacje na skali porządkowej (odległość GDM2). Przedstawiono odpowiednie formuły odległości oraz omówiono ich zalety i wady. Agregacji preferencji za pomocą poszczególnych miar odległości dokonano z wykorzystaniem programu R.