STATISTICS IN TRANSITION new series, September 2017
Vol. 18, No. 3, pp. 521–540, DOI: 10.21307/stattrans-2016-084

SELECTING THE OPTIMAL MULTIDIMENSIONAL SCALING PROCEDURE FOR METRIC DATA WITH R ENVIRONMENT

Marek Walesiak[1], **Andrzej Dudek**[2]

ABSTRACT

In multidimensional scaling (MDS) carried out on the basis of a metric data matrix (interval, ratio), the main decision problems relate to the selection of the method of normalization of the values of the variables, the selection of distance measure and the selection of MDS model. The article proposes a solution that allows choosing the optimal multidimensional scaling procedure according to the normalization methods, distance measures and MDS model applied. The study includes 18 normalization methods, 5 distance measures and 3 types of MDS models (ratio, interval and spline). It uses two criteria for selecting the optimal multidimensional scaling procedure: Kruskal's *Stress*-1 fit measure and Hirschman-Herfindahl *HHI* index calculated based on Stress per point values. The results are illustrated by an empirical example.

Key words: multidimensional scaling, normalization of variables, distance measures, *HHI* index, R program.

1. Introduction

Multidimensional scaling is a method that represents (dis)similarity data as distances in a low-dimensional space (typically 2 or 3 dimensional) in order to make these data accessible to visual inspection and exploration (Borg, Groenen, 2005, p. 3). The dimensions are not directly observable. They have the nature of latent variables. MDS allows the similarities and differences between the analyzed objects to be explained.

Multidimensional scaling is a widely used technique in many areas, including psychology (Takane, 2007), sociology (Pinkley, Gelfand, Duan, 2005), linguistics

[1] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra. The project is financed by the Polish National Science Centre, decision DEC-2015/17/B/HS4/00905. E-mail: marek.walesiak@ue.wroc.pl.
[2] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra. E-mail: andrzej.dudek@ue.wroc.pl.

(Embleton, Uritescu, Wheeler, 2013), marketing research (Cooper, 1983), tourism (Marcussen, 2014) and geography (Golledge, Ruhton, 1972).

The starting point of multidimensional scaling is a distance matrix (dissimilarities) between objects in *m*-dimensional space $\boldsymbol{\delta} = [\delta_{ik}]$, where $i, k = 1, \ldots, n$ is the number of the object. Methods of determining the distance matrix $\boldsymbol{\delta} = [\delta_{ik}]$ can be divided into direct (typically result from similarity ratings on object pairs, from rankings, or from card-sorting tasks) and indirect (they can be derived from other data) methods (see, e.g. Borg, Groenen, 2005, pp. 111-133).

The article uses an indirect method in which the starting point is a metric data matrix $\mathbf{X} = [x_{ij}]$ (x_{ij} – the value of the *j*-th variable for the *i*-th object, $j = 1, \ldots, m$ – the number of metric variable), for which observations are obtained from secondary data sources. It is a typical situation in socio-economic research.

The normalization of variables is carried out when the variables describing the analyzed objects are measured on metric scales (interval or ratio). The characteristics of measurement scales were discussed, e.g. in the study by (Stevens, 1946). The purpose of normalization is to achieve the comparability of variables.

Metric data that requires normalization of variables complicates the problem of choosing a multidimensional scaling procedure. The article proposes a solution that allows the choice of the optimal multidimensional scaling procedure, carried out on the basis of metric data (interval, ratio), according to the normalization methods, distance measures and MDS model applied. The study included 18 normalization methods, 5 distance measures and MDS models (ratio, interval and spline – e.g. polynomial function of second or third degree). For instance, ten normalization methods, five distance measures and four MDS models give 200 multidimensional scaling procedures.

The authors of the monograph (Borg, Groenen, Mair, 2013, chapter 7) pointed out the typical mistakes made by users of multidimensional scaling. A frequent mistake on the part of users of MDS results is to evaluate Stress mechanically (rejecting an MDS solution because its Stress seems "too high"). In their opinion (Borg, Groenen, Mair, 2013, p. 68) "An MDS solution can be robust and replicable, even if its Stress value is high" and "Stress, moreover, is a *summative* index for *all* proximities. It does not inform the user how well a *particular* proximity value is represented in the given MDS space". In addition, we should take into account Stress per point measure (the average of the squared error terms for each point) and acceptability of MDS results (based on "Shepard diagram").

To solve the problem of choosing the optimal multidimensional scaling procedure, two criteria were applied: Kruskal's *Stress*-1 (*Stress* – Standardized residual sum of squares) fit measure and the Hirschman-Herfindahl *HHI* index, calculated based on Stress per point values (*spp*). The article proposes an algorithm that allows the selection of the optimal multidimensional scaling procedure with implementation in `mdsOpt` package of R program (Walesiak, Dudek, 2017b).

The results are illustrated by an empirical example.

2. Multidimensional scaling based on metric data

A general scheme of multidimensional scaling performed on metric data is as follows:

$$P \rightarrow A \rightarrow X \rightarrow \mathbf{X} \rightarrow \mathbf{Z} \rightarrow \boldsymbol{\delta} \rightarrow S \rightarrow \mathbf{d} \rightarrow \mathbf{V} \rightarrow I \,, \qquad (1)$$

where:

P – choice of research problem,

A – selection of objects,

X – selection of variables,

$\mathbf{X}$ – collecting data and construction of data matrix $\mathbf{X} = [x_{ij}]_{nxm}$ for $i,k = 1,...,n$ and $j = 1,...,m$ (x_{ij} – the value of the *j*-th variable for the *i*-th object),

$\mathbf{Z}$ – choice of variable normalization method and construction of normalized data matrix $\mathbf{Z} = [z_{ij}]_{nxm}$ for $i,k = 1,...,n$ and $j = 1,...,m$ (z_{ij} – the normalized value of the *j*-th variable for the *i*-th object),

$\boldsymbol{\delta}$ – selection of distance measure (see Table 3) and construction of distance matrix in *m*-dimensional space $\boldsymbol{\delta} = [\delta_{ik}(\mathbf{Z})]_{nxn}$ for $i,k = 1,...,n$,

S – perform multidimensional scaling (MDS): $f : \delta_{ik}(\mathbf{Z}) \rightarrow d_{ik}(\mathbf{V})$ for all pairs (i,k) – mapping distances in *m*-dimensional space $\delta_{ik}(\mathbf{Z})$ into corresponding distances $d_{ik}(\mathbf{V})$ in *q*-dimensional space ($q < m$) by a representation function *f*. The distances $d_{ik}(\mathbf{V})$ are always unknown, i.e. MDS must find a configuration $\mathbf{V}$ of predetermined dimensions *q* on which the distances are computed,

$\mathbf{d}$ – Euclidean distance matrix in *q*-dimensional space ($q < m$, typically *q* equals 2 or 3) $\mathbf{d} = [d_{ik}(\mathbf{V})]_{nxn}$ for $i,k = 1,...,n$,

$\mathbf{V}$ – configuration of objects in *q*-dimensional space $\mathbf{V} = [v_{ij}]_{nxq}$,

I – interpretation of multidimensional scaling results in *q*-dimensional space.

In SMACOF (Scaling by Majorizing a Complicated Function) algorithm we minimize Stress (2) over the configuration matrix $\mathbf{V}$ by an iterative procedure (see Borg, Groenen, 2005, pp. 204-205):

1. Set $\mathbf{V} = \mathbf{V}^{[0]}$, where $\mathbf{V}^{[0]}$ is some nonrandom or random start configuration. Starting solution is usually Torgerson-Gower classical scaling (Torgerson, 1952; Gower, 1966). Set iteration counter $k = 0$. Set ε to a small positive constant (convergence criterion), i.e. $\varepsilon = 0.000001$.
2. Find optimal disparities $\hat{d}_{ik}$ for fixed distances $d_{ik}(\mathbf{V}^{[0]})$.
3. Standardize (to avoid degenerated solution) $\hat{d}_{ik}$ so that $\eta_{\hat{d}}^2 = n(n-1)/2$.
4. Compute Stress function $\sigma_r^{[0]} = \sigma_r(\hat{\mathbf{d}}, \mathbf{V}^{[0]})$:

$$\begin{aligned}\sigma_r(\hat{\mathbf{d}}, \mathbf{V}) &= \sum_{i<k} w_{ik}(d_{ik}(\mathbf{V}) - \hat{d}_{ik})^2 \\ &= \sum_{i<k} w_{ik}\hat{d}_{ik}^2 + \sum_{i<k} w_{ik} d_{ik}^2(\mathbf{V}) - 2\sum_{i<k} w_{ik}\hat{d}_{ik} d_{ik}(\mathbf{V}) \\ &= \eta_{\hat{d}}^2 + \eta^2(\mathbf{V}) - 2\rho(\hat{\mathbf{d}}, \mathbf{V}). \qquad (2)\end{aligned}$$

where: $\hat{d}_{ik}$ – d-hats, disparities, target distances or pseudo distances (see Borg, Groenen 2005, p. 199). $\hat{d}_{ik} = f(\delta_{ik})$ by defining f in different ways: $\hat{d}_{ik} = b \cdot \delta_{ik}$ – ratio MDS; $\hat{d}_{ik} = a + b \cdot \delta_{ik}$ – interval MDS, $\hat{d}_{ik} = a + b \cdot \delta_{ik} + c \cdot \delta_{ik}^2$ – spline MDS (polynomial function of second degree);
$w_{ik} = 1$ – for object pair i, k a dissimilarity has been observed, $w_{ik} = 0$ – otherwise.
Set $\sigma_r^{[-1]} = \sigma_r^{[0]}$.

5. While $k = 0$ or ($\sigma_r^{[k-1]} - \sigma_r^{[k]} > \varepsilon$ and $k \leq$ maximum iterations) do
6. Increase iteration number k by one ($k := k + 1$).
7. Compute Guttman transform $\mathbf{V}^{[k]}$ (see Borg, Groenen, 2005, p. 191; De Leeuw, Mair, 2009, p. 5).
8. Find optimal disparities $\hat{d}_{ik}$ for fixed distances $d_{ik}(\mathbf{V}^{[k]})$.
9. Standardize $\hat{d}_{ik}$ so that $\eta_{\hat{d}}^2 = n(n-1)/2$.
10. Compute $\sigma_r^{[k]} = \sigma_r(\hat{\mathbf{d}}, \mathbf{V}^{[k]})$.
11. Set $\mathbf{V} = \mathbf{V}^{[k]}$,
12. End while.

A flowchart of the SMACOF algorithm is given in Figure 1.

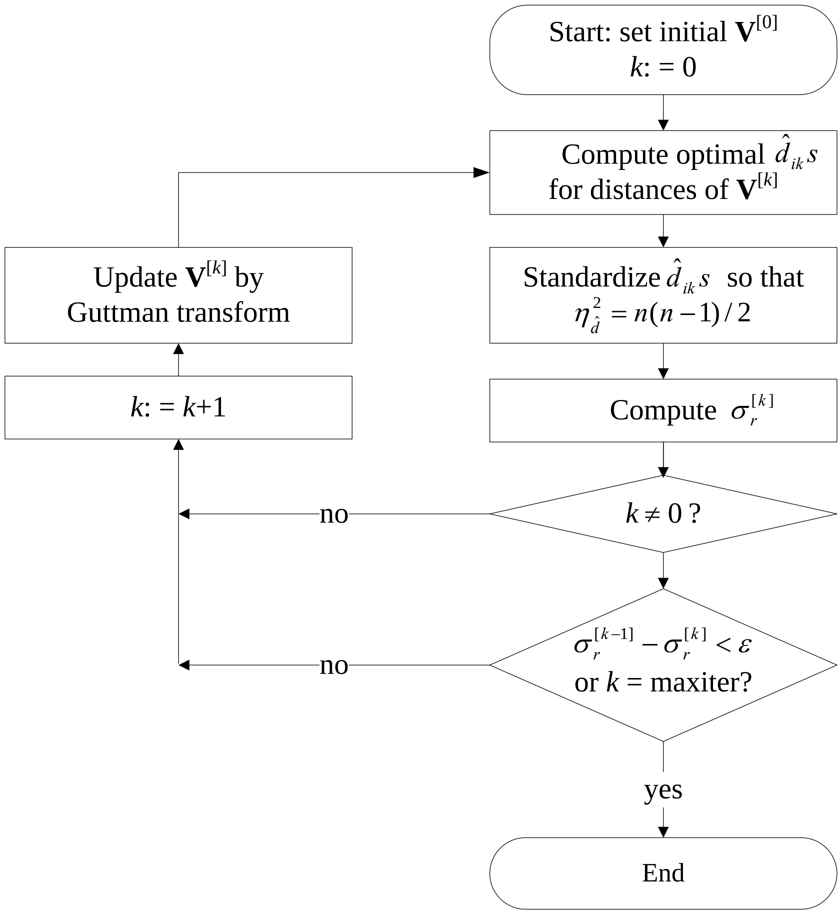

Figure 1. The flowchart of the majorization algorithm (SMACOF)

Source: Borg, Groenen, 2005, p. 205.

In other multidimensional scaling algorithms, different fit measures are applied (see, e.g. Borg, Groenen, 2005, pp. 250-254): Kruskal's *Stress*-1, Kruskal and Carroll *Stress*-2, the Guttman-Lingoes coefficient of alienation, *S*-*Stress* of Takane, Young and De Leeuw.

3. Criteria for the selection of the optimal multidimensional scaling procedure

The article proposes a solution that allows the optimal multidimensional scaling procedure to be chosen. The study uses the function `smacofSym` of `smacof` package od R program (R Development Core Team, 2017). In the function `smacofSym` of `smacof` package (Mair et al., 2017) basic decision problems involve the following selection:

– normalization method (the analysis included 18 normalization methods),
– distance measure (the analysis included 5 distance measures),
– MDS model (the analysis included: ratio MDS, interval MDS, spline MDS).

Table 1 presents normalization methods, given by linear formula (3), which were used in the selection of the optimal MDS procedure (see Jajuga, Walesiak, 2000, pp. 106-107; Zeliaś, 2002, p. 792):

$$z_{ij} = b_j x_{ij} + a_j = \frac{x_{ij} - A_j}{B_j} = \frac{1}{B_j} x_{ij} - \frac{A_j}{B_j} \quad (b_j > 0), \tag{3}$$

where: x_{ij} – the value of *j*-th variable for the *i*-th object,

z_{ij} – the normalized value of *j*-th variable for the *i*-th object,

A_j – shift parameter to arbitrary zero for the *j*-th variable,

B_j – scale parameter for the *j*-th variable,

$a_j = -A_j / B_j$, $b_j = 1/B_j$ – parameters for the *j*-th variable presented in Table 1.

Table 1. Normalization methods

Type	Method	Parameter		Scale of variables	
		b_j	a_j	BN	AN
n1	Standardization	$1/s_j$	$-\bar{x}_j / s_j$	ratio or interval	interval
n2	Positional standardization	$1/mad_j$	$-med_j / mad_j$	ratio or interval	interval
n3	Unitization	$1/r_j$	$-\bar{x}_j / r_j$	ratio or interval	interval
n3a	Positional unitization	$1/r_j$	$-med_j / r_j$	ratio or interval	interval
n4	Unitization with zero minimum	$1/r_j$	$-\min_i\{x_{ij}\} / r_j$	ratio or interval	interval
n5	Normalization in range [–1; 1]	$\frac{1}{\max_i \lvert x_{ij} - \bar{x}_j \rvert}$	$\frac{-\bar{x}_j}{\max_i \lvert x_{ij} - \bar{x}_j \rvert}$	ratio or interval	interval
n5a	Positional normalization in range [–1; 1]	$\frac{1}{\max_i \lvert x_{ij} - med_j \rvert}$	$\frac{-med_j}{\max_i \lvert x_{ij} - med_j \rvert}$	ratio or interval	interval

Table 1. Normalization methods (cont.)

Type	Method	Parameter		Scale of variables	
		b_j	a_j	BN	AN
n6	Quotient transformations	$1/s_j$	0	ratio	ratio
n6a		$1/mad_j$	0	ratio	ratio
n7		$1/r_j$	0	ratio	ratio
n8		$1/\max_i\{x_{ij}\}$	0	ratio	ratio
n9		$1/\bar{x}_j$	0	ratio	ratio
n9a		$1/med_j$	0	ratio	ratio
n10		$1/\sum_{i=1}^{n} x_{ij}$	0	ratio	ratio
n11		$1/\sqrt{\sum_{i=1}^{n} x_{ij}^2}$	0	ratio	ratio
n12	Normalization	$\frac{1}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$	$\frac{-\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$	ratio or interval	interval
n12a	Positional normalization	$\frac{1}{\sqrt{\sum_{i=1}^{n}(x_{ij}-med_j)^2}}$	$\frac{-med_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-med_j)^2}}$	ratio or interval	interval
n13	Normalization with zero being the central point	$\frac{1}{r_j/2}$	$-\frac{m_j}{r_j/2}$	ratio or interval	interval

BN – before normalization, AN – after normalization, $\bar{x}_j$ – mean for the *j*-th variable, s_j – standard deviation for the *j*-th variable, r_j – range for the *j*-th variable, $m_j = \frac{\max_i\{x_{ij}\} + \min_i\{x_{ij}\}}{2}$ – mid-range for the *j*-th variable, $med_j = \underset{i}{med}(x_{ij})$ – median for the *j*-th variable, $mad_j = \underset{i}{mad}(x_{ij})$ – median absolute deviation for the *j*-th variable.

Source: Based on (Jajuga, Walesiak, 2000; Walesiak, Dudek, 2017a).

Column 1 in Table 1 presents the type of normalization method adopted as the function `data.Normalization` of `clusterSim` package (Walesiak, Dudek, 2017a). Similar procedure for data normalization is available as the function `scale` of `base` package. In this function the researcher defines the parameters A_j and B_j.

Due to the fact that the groups of A, B, C and D (see Table 2) normalization methods give identical multidimensional scaling results, further analysis covers the first methods of the identified groups (n1, n2, n3, n9), as well as the other methods (n5, n5a, n8, n9a, n11, n12a).

Table 2. The groups of normalization methods resulting in identical distance matrices

Groups of normalization methods	Normalization methods	
	GDM1 distance	Minkowski distances, squared Euclidean distance*
A	n1, n6, n12	n1, n6, n12
B	n2, n6a	n2, n6a
C	n3, n3a, n4, n7, n13	n3, n3a, n4, n7, n13
D	n9, n10	n9, n10

* after dividing distances in each distance matrix by the maximum value.

Source: Own presentation.

Table 3 presents selected distance measures for metric data that have been used in the selection of the optimal multidimensional scaling procedure.

Distance GDM1 is available as a function of `dist.GDM` of `clusterSim` package (Walesiak, Dudek, 2017) and the remaining distances in Table 3 are available in the function `dist` of `stats` package (R Development Core Team, 2017).

The initial point of the application of `smacofSym` function is to determine the following values of arguments:

– convergence criterion (eps=1e-06),
– maximum number of iterations (itmax=1000).

These parameters can be changed by the user.

The selection of the optimal procedure for multidimensional scaling takes place in several stages:

1. Set the number of dimensions in MDS to two (ndim=2).
2. Taking into account in the analysis 10 normalization methods, 5 distance measures and 2 MDS models, there are 100 multidimensional scaling procedures. Multidimensional scaling is performed for each procedure separately. It then orders the procedures by increasing *Stress*-1 fit measure (see e.g. Borg, Groenen, Mair, 2013, p. 23):

$$Stress\text{-}1_p = \sqrt{\sum_{i<k}[d_{ik}(\mathbf{V}) - \hat{d}_{ik}]^2 \Big/ \sum_{i<k} d_{ik}^2(\mathbf{V})}\,, \tag{4}$$

where: $p = 1,\ldots,100$ – multidimensional scaling procedure number.

Table 3. Distance measures for metric (interval, ratio) data

Name	Distance δ_{ik}	Range	Allowed normalization
Minkowski $(p \geq 1)$	$\sqrt[p]{\sum_{j=1}^{m} \lvert z_{ij} - z_{kj} \rvert^p}$	$[0; \infty)$	n1-n13
Manhattan $(p = 1)$	$\sum_{j=1}^{m} \lvert z_{ij} - z_{kj} \rvert$	$[0; \infty)$	n1-n13
Euclidean $(p = 2)$	$\sqrt{\sum_{j=1}^{m} \left(z_{ij} - z_{kj}\right)^2}$	$[0; \infty)$	n1-n13
Chebyshev (maximum) ($p \to \infty$)	$\max_j \lvert z_{ij} - z_{kj} \rvert$	$[0; \infty)$	n1-n13
Squared Euclidean	$\sum_{j=1}^{m} \left(z_{ij} - z_{kj}\right)^2$	$[0; \infty)$	n1-n13
GDM1	$\frac{1}{2} - \frac{\sum_{j=1}^{m}(z_{ij} - z_{kj})(z_{kj} - z_{ij}) + \sum_{j=1}^{m}\sum_{\substack{l=1 \\ l \neq i,k}}^{n}(z_{ij} - z_{lj})(z_{kj} - z_{lj})}{2\left[\sum_{j=1}^{m}\sum_{l=1}^{n}(z_{ij} - z_{lj})^2 \cdot \sum_{j=1}^{m}\sum_{l=1}^{n}(z_{kj} - z_{lj})^2\right]^{\frac{1}{2}}}$	$[0; 1]$	n1-n13

$i, k, l = 1, \ldots, n$ – object number, m – the number of objects, $j = 1, \ldots, m$ – variable number, m – the number of variables, $z_{ij} (z_{kj}, z_{lj})$ – the normalized value of the *j*-th variable for the *i*-th (*k*-th, *l*-th) object.

Source: Based on (Everitt et al., 2011, pp. 49-50; Jajuga, Walesiak, Bąk, 2003).

3. Based on Stress per point (*spp*) values (Stress contribution in percentages), the Hirschman-Herfindahl index is calculated (Herfindahl, 1950; Hirschman, 1964):

$$HHI_p = \sum_{i=1}^{n} spp_{pi}^2 , \tag{5}$$

where: $i = 1, \ldots, n$ – object number.

The HHI_p index takes values in the interval $\left[\frac{10{,}000}{n}; 10{,}000\right]$. The value $\frac{10{,}000}{n}$ means that the distribution of errors for individual objects is uniform ($\underset{i}{\forall} spp_i = \frac{100}{n}$). The maximal value appears when summary fit measure (*Stress*-1) is the result of loss assigned only to one object. For other objects, loss function will be equal to zero. The optimal situation for a multidimensional scaling procedure is the minimal value of the HHI_p index.

4. The chart with $Stress\text{-}1_p$ fit measure value on *x*-axis and HHI_p index on *y*-axis for *p* procedures of multidimensional scaling is drawn.

5. The maximal acceptable value of $Stress\text{-}1$ is assumed as s. For all multidimensional scaling procedures for which $Stress\text{-}1_p \le s$, we chose the one for which $\min_p \{HHI_p\}$ occurs.
6. Multidimensional scaling for the selected procedure is performed along with checkout that in the sense of interpretation results are acceptable. Based on the Shepard diagram, the correctness of the model scaling will be evaluated. If the results are acceptable the procedure ends, otherwise it returns to step 1 and multidimensional scaling for three dimensions is performed (ndim=3).

4. Empirical results

The empirical study uses the statistical data presented in the article (Gryszel, Walesiak, 2014) and referring to the attractiveness level of 29 Lower Silesian counties. The evaluation of tourist attractiveness of Lower Silesian counties was performed using 16 metric variables (measured on a ratio scale):

x1 – beds in hotels per 1 km^2 of a county area,

x2 – number of nights spent daily by resident tourists (Poles) per 1,000 inhabitants of a county,

x3 – number of nights spent daily by foreign tourists per 1,000 inhabitants of a county,

x4 – gas pollution emission in tons per 1 km^2 of a county area,

x5 – number of criminal offences and crimes against life and health per 1,000 inhabitants of a county,

x6 – number of property crimes per 1,000 inhabitants of a county,

x7 – number of historical buildings per 100 km^2 of a county area,

x8 – % of a county forest cover,

x9 – % share of legally protected areas within a county area,

x10 – number of events as well as cultural and tourist ventures in a county,

x11 – number of natural monuments calculated per 1 km^2 of a county area,

x12 – number of tourist economy entities per 1,000 inhabitants of a county (natural and legal persons),

x13 – expenditure of municipalities and counties on tourism, culture and national heritage protection as well as physical culture per 1 inhabitant of a county in Polish zlotys (PLN),

x14 – cinema attendance per 1,000 inhabitants of a county,

x15 – museum visitors per 1,000 inhabitants of a county,

x16 – number of construction permits (hotels and accommodation buildings, commercial and service buildings, transport and communication buildings, civil and water engineering constructions) issued in a county in the years 2011-2012, per 1 km^2 of a county area.

The statistical data were collected in 2012 and come from the Local Data Bank of the Central Statistical Office of Poland; the data for x7 variable only were obtained from the regional conservation officer.

Variables (x4, x5 and x6) take the form of destimulants, x9 is a nominant (50% level was adopted as the optimal one). The other variables represent stimulants, whereas x9 nominant was transformed into a stimulant. The definitions of stimulants, destimulants and nominants are available in the study, e.g. (Walesiak, 2016).

A pattern object and an anti-pattern object were added to the set of 29 counties (see Walesiak, 2016). Therefore, the data matrix covers 31 objects described by 16 variables. The coordinates of a pattern object cover the most preferred preference variable (stimulants, destimulants and nominants) values. The coordinates of an anti-pattern object cover the least preferred preference variable values.

The article uses its own script of package `mdsOpt` of R program (Walesiak, Dudek, 2017b) to choose the optimal procedure for multidimensional scaling due to normalization methods, selected distance measures and MDS models (developed in accordance with the methodology described in section 3).

The measurement of variables on a ratio scale accepts all normalization methods (hence the study covered 18 methods). Due to the fact that the groups of A, B, C and D normalization methods give identical multidimensional scaling results (see Table 2), further analysis covers the first methods of the identified groups (n1, n2, n3, n9), as well as the other methods (n5, n5a, n8, n9a, n11, n12a).

Ordering results of 100 multidimensional scaling procedures (10 normalization methods x 5 distance measures x 2 MDS models) according to formula (4) are presented in Table 4. In addition, Table 4 shows values of HHI_p index for each MDS procedure.

Table 4. Ordering results of 100 multidimensional scaling procedures

p	nm	MDS model	Distance measure	*Stress*-1	*HHI*	*p*	nm	MDS model	Distance measure	*Stress*-1	*HHI*
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
1	n9a	interval	euclidean	0.0311	844	51	n2	ratio	seuclidean	0.1391	1328
2	n2	interval	euclidean	0.0369	685	52	n11	ratio	GDM1	0.1391	495
3	n9a	ratio	euclidean	0.0404	715	53	n5a	interval	seuclidean	0.1400	663
4	n9a	interval	maximum	0.0408	1276	54	n5	ratio	seuclidean	0.1402	797
5	n9a	ratio	maximum	0.0441	1230	55	n5a	interval	euclidean	0.1405	508
6	n2	interval	maximum	0.0505	908	56	n11	ratio	manhattan	0.1414	453
7	n2	ratio	euclidean	0.0546	520	57	n5a	ratio	seuclidean	0.1436	791
8	n2	ratio	maximum	0.0576	794	58	n9	ratio	euclidean	0.1473	464
9	n9a	interval	manhattan	0.0627	867	59	n9a	ratio	seuclidean	0.1478	1289
10	n9a	ratio	manhattan	0.0687	645	60	n8	ratio	manhattan	0.1483	428
11	n2	interval	manhattan	0.0704	755	61	n3	ratio	manhattan	0.1502	419
12	n2	interval	GDM1	0.0770	605	62	n1	ratio	manhattan	0.1530	410
13	n9a	interval	GDM1	0.0793	593	63	n5	ratio	manhattan	0.1531	421

Table 4. Ordering results of 100 multidimensional scaling procedures (cont.)

p	nm	MDS model	Distance measure	*Stress*-1	*HHI*	*p*	nm	MDS model	Distance measure	*Stress*-1	*HHI*
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
14	n2	ratio	manhattan	0.0839	521	64	n12a	ratio	manhattan	0.1543	409
15	n2	ratio	GDM1	0.0894	887	65	n5a	ratio	manhattan	0.1548	422
16	n9a	ratio	GDM1	0.0969	924	66	n8	interval	GDM1	0.1598	486
17	n9	interval	manhattan	0.0985	577	67	n8	ratio	GDM1	0.1608	489
18	n9	interval	euclidean	0.1056	580	68	n9	interval	maximum	0.1610	554
19	n9	interval	seuclidean	0.1087	813	69	n3	interval	GDM1	0.1640	473
20	n11	interval	manhattan	0.1092	500	70	n3	ratio	GDM1	0.1653	476
21	n8	interval	manhattan	0.1149	476	71	n1	interval	GDM1	0.1677	431
22	n11	interval	seuclidean	0.1149	739	72	n1	ratio	GDM1	0.1691	435
23	n3	interval	manhattan	0.1155	469	73	n11	ratio	euclidean	0.1698	427
24	n2	interval	seuclidean	0.1161	865	74	n12a	interval	GDM1	0.1718	430
25	n9	ratio	seuclidean	0.1164	1102	75	n12a	ratio	GDM1	0.1732	434
26	n9	interval	GDM1	0.1166	545	76	n5	interval	GDM1	0.1737	494
27	n9	ratio	GDM1	0.1166	545	77	n5	ratio	GDM1	0.1738	494
28	n11	interval	euclidean	0.1168	497	78	n5a	interval	GDM1	0.1774	493
29	n11	ratio	seuclidean	0.1179	922	79	n5a	ratio	GDM1	0.1774	493
30	n1	interval	manhattan	0.1186	457	80	n11	interval	maximum	0.1874	494
31	n12a	interval	manhattan	0.1199	455	81	n9	ratio	maximum	0.1878	489
32	n9a	interval	seuclidean	0.1204	791	82	n8	ratio	euclidean	0.1883	419
33	n5	interval	manhattan	0.1207	479	83	n1	ratio	euclidean	0.1908	399
34	n5a	interval	manhattan	0.1225	479	84	n5	ratio	euclidean	0.1914	420
35	n8	interval	seuclidean	0.1255	688	85	n3	ratio	euclidean	0.1921	411
36	n9	ratio	manhattan	0.1257	486	86	n12a	ratio	euclidean	0.1923	398
37	n3	interval	seuclidean	0.1263	694	87	n5a	ratio	euclidean	0.1925	418
38	n8	ratio	seuclidean	0.1274	803	88	n1	interval	maximum	0.2229	437
39	n3	ratio	seuclidean	0.1279	802	89	n12a	interval	maximum	0.2242	441
40	n1	interval	seuclidean	0.1280	719	90	n11	ratio	maximum	0.2260	442
41	n8	interval	euclidean	0.1292	474	91	n8	interval	maximum	0.2307	460
42	n1	ratio	seuclidean	0.1297	845	92	n5a	interval	maximum	0.2368	424
43	n12a	interval	seuclidean	0.1300	718	93	n3	interval	maximum	0.2398	463
44	n1	interval	euclidean	0.1303	421	94	n5	interval	maximum	0.2442	443
45	n3	interval	euclidean	0.1307	461	95	n1	ratio	maximum	0.2547	396
46	n12a	ratio	seuclidean	0.1318	845	96	n12a	ratio	maximum	0.2557	395
47	n12a	interval	euclidean	0.1322	421	97	n5a	ratio	maximum	0.2606	394
48	n5	interval	seuclidean	0.1369	666	98	n8	ratio	maximum	0.2618	414
49	n11	interval	GDM1	0.1381	493	99	n3	ratio	maximum	0.2652	418
50	n5	interval	euclidean	0.1382	500	100	n5	ratio	maximum	0.2667	405

nm – normalization method; seuclidean – squared Euclidean distance.

Source: Authors' compilation using mdsOpt package and R program.

In the conducted study the maximal acceptable value of $Stress\text{-}1_p$ fit measure has been set to 0.15. Figure 2 presents the chart with $Stress\text{-}1_p$ fit measure value on *x*-axis and HHI_p index on *y*-axis for *p* procedures of multidimensional scaling.

Among acceptable multidimensional scaling procedures, for which $Stress\text{-}1_p \leq 0.15$, we chose the one for each occurs $\min_p \{HHI_p\}$ has been chosen. It is the procedure 47: `n12a` normalization method (positional normalization), interval MDS model, Euclidean distance.

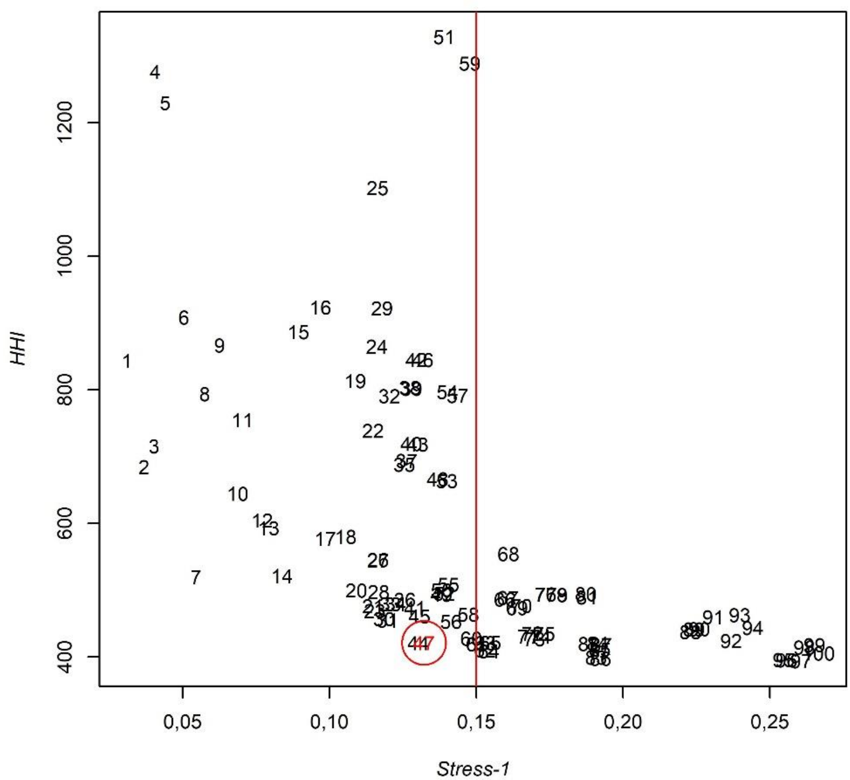

Figure 2. The values of $Stress\text{-}1_p$ fit measure and HHI_p index for *p* multidimensional scaling procedures

Source: Authors' compilation using `mdsOpt` *package of R program.*

The results of multidimensional scaling (procedure 47) of 31 objects (29 Lover Silesian counties, pattern and anti-pattern object) according to the level of tourist attractiveness are presented on Figure 3.

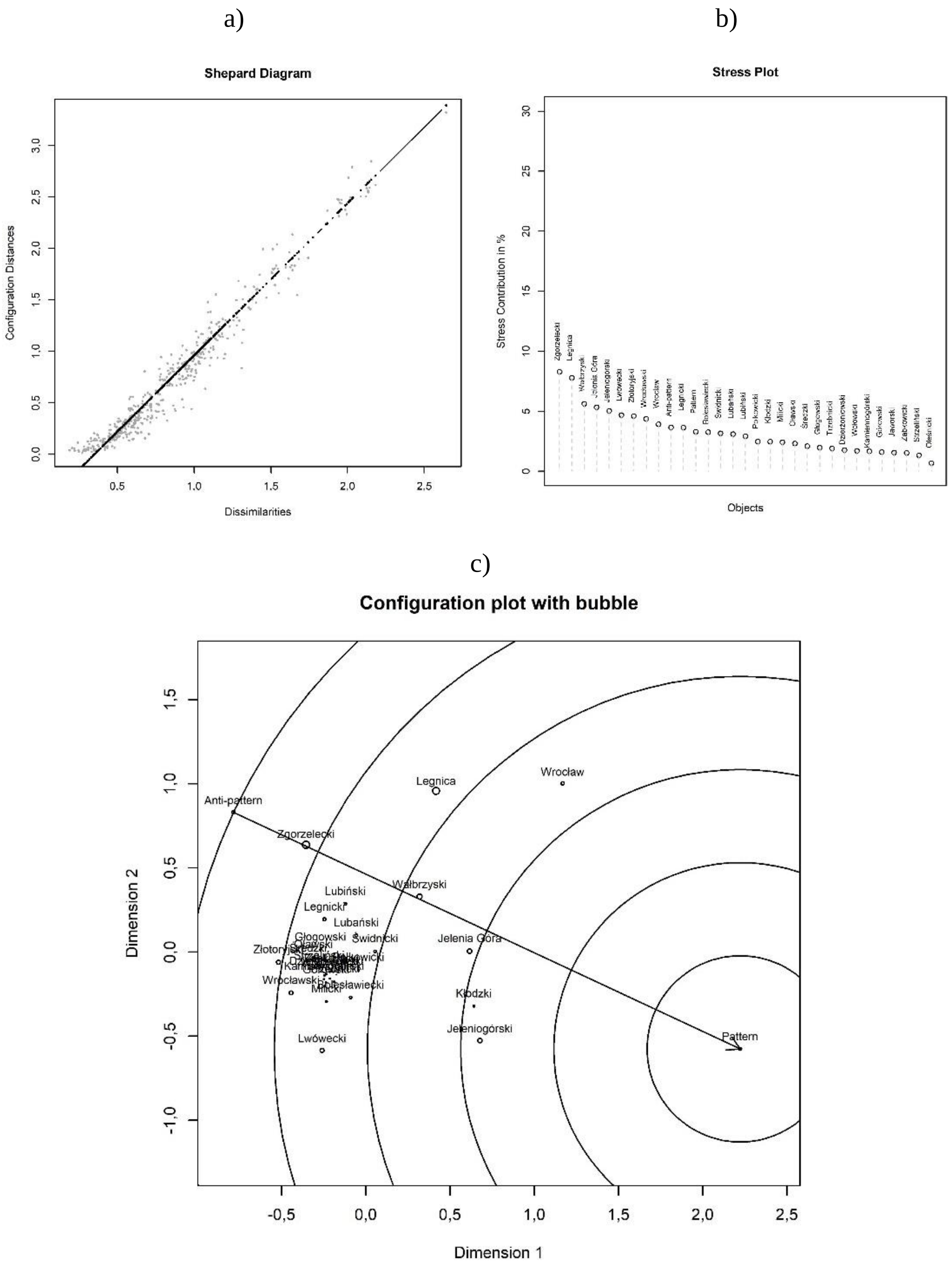

Figure 3. The results of multidimensional scaling (procedure 47) of 31 objects (29 Lover Silesian counties, pattern and anti-pattern) according to the level of tourist attractiveness (d_{ik} – Configuration Distances, δ_{ik} – Dissimilarities)

Source: Authors' compilation using R program.

Figure 3c (Configuration plot with bubble) presents additional quota of each object in total error is shown by the size of radius of the circle around each object. Shepard diagram (Figure 3a) confirms the correctness of the chosen scaling model (Pearson correlation coefficient $r = 0.9794$). Figure 3c (Configuration plot with bubble) shows the axis of the set, which is the shortest connection between the pattern and anti-pattern of development. It indicates the level of development of the tourist attractiveness of counties. Objects that are closer to the pattern of development have higher levels of tourist attractiveness. The isoquants[3] of development (curves of similar development) have been established from the point indicating pattern object. Figure 3c shows six isoquants. The same level of development may be achieved by objects from different locations on the same isoquant of development (due to different configuration of values of variables).

As opposed to the best MDS procedure (47) we show the results for one of the worst procedures (4): `n9a` normalization method, interval MDS model, maximum (Chebyshev) distance. Overall Stress for procedure 4 (0.0408) is significantly better than for procedure 47 (0.1322). The results of multidimensional scaling for procedure 4 according to the level of tourist attractiveness are presented in Figure 4.

Figure 4b (Stress Plot) indicates that objects Jeleniogórski, Anti-pattern and Zgorzelecki contribute most to the overall Stress (55.6%). It also shows (see Shepard diagram – in the lower left-hand corner) that two points (distance between Jeleniogórski county and Anti-pattern object; Jeleniogórski county and Zgorzelecki county) are outliers. These outliers contribute over-proportionally to the total Stress. MDS configuration (Figure 4c) does not represent all proximities equally well. Jeleniogórski county is one of the best of Lover Silesian counties in terms of the level of tourist attractiveness. In Figure 4c (Configuration plot with bubble) this county lies near Anti-pattern object (the worst object). The greater the value of the HHI_p index, the worse is the effect of multidimensional scaling in terms of representing real relationships between objects.

5. Summary and limitations of presented proposal

The article proposes a methodology that allows the selection of the optimum procedure due to the used methods of normalization, distance measures and scaling model of multidimensional scaling carried out on the basis of the metric data matrix. The study includes 18 methods of normalization, 5 distance measures and 3 models of scaling (ratio, interval and spline scaling).

Own package `mdsOpt` of R program to choose the optimal procedure for multidimensional scaling due to the normalization methods of variable values, distance measures and scaling models has been developed. On the basis of the proposed methodology research results are illustrated by an empirical example with the use of the function `smacofSym` of `smacof` package in order to find the

[3] Isoquants were illustrated using `draw.circle` function of `plotrix` package (Lemon et al., 2017).

optimal procedure for multidimensional scaling of set of objects representing 29 counties in Lower Silesia according to the level of tourist attractiveness.

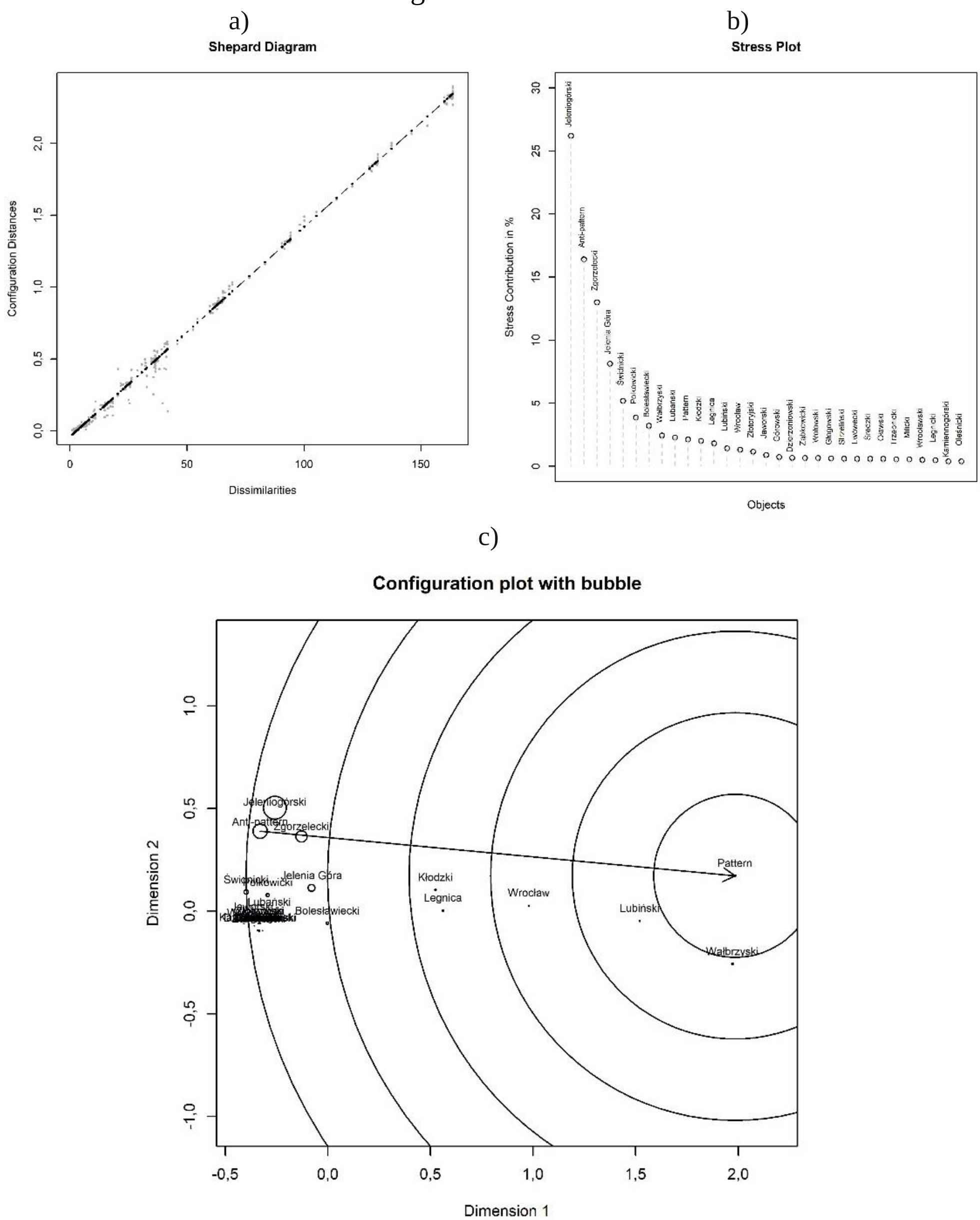

Figure 4. The results of multidimensional scaling (procedure 4) of 31 objects (29 Lover Silesian counties, pattern and anti-pattern) according to the level of tourist attractiveness

Source: Authors' compilation using R program.

The proposed methodology uses two criteria for selecting the optimal procedure for multidimensional scaling: *Stress*-1 loss function and the value of the Hirschman-Herfindahl *HHI* index calculated on the basis of the decomposition *Stress*-1 error by objects.

In step 5 the maximal acceptable value of fit measure $Stress\text{-}1 = s$ has been arbitrary assumed. The extent to which error distribution for each object may deviate from the uniform distribution is not determined. Among the procedures of multidimensional scaling for which $Stress\text{-}1_p \leq s$, the one for which $\min_p \{HHI_p\}$ occurs is selected. This constraint does not essentially limit the presented proposal as the additional criteria for acceptability of the results of multidimensional scaling plots, such as "Shepard diagram" and "Residual plot", make it possible to evaluate the fit quality of the chosen scaling model, and to identify outliers (De Leeuw, Mair, 2015).

REFERENCES

BORG, I., GROENEN, P. J. F., (2005). Modern Multidimensional Scaling. Theory and Applications, 2nd Edition, Springer Science+Business Media, New York. ISBN: 978-0387-25150-9, URL http://www.springeronline.com/0-387-25150-2.

BORG, I., GROENEN, P. J. F., MAIR, P., (2013). Applied Multidimensional Scaling, Springer, Heidelberg, New York, Dordrecht, London, URL http://dx.doi.org/10.1007/978-3-642-31848-1.

COOPER, L. G., (1983). A review of multidimensional scaling in marketing research, Applied Psychological Measurement, Vol. 7, No. 4, pp. 427–450, URL https://doi.org/10.1177/014662168300700404.

DE LEEUW, J., MAIR, P., (2009). Multidimensional Scaling Using Majorization: SMACOF in R. Journal of Statistical Software, 31 (3), pp. 1–30, URL http://dx.doi.org/10.18637/jss.v031.i03.

DE LEEUW, J., MAIR, P., (2015). Shepard Diagram, Wiley StatsRef: Statistics Reference Online, Wiley, URL http://dx.doi.org/10.1002/9781118445112.stat06268.pub2.

EMBLETON, S., URITESCU, D., WHEELER, E. S., (2013). Defining dialect regions with interpretations: Advancing the multidimensional scaling approach, Literary and Linguistic Computing, Vol. 28, No. 1, pp. 13–22, URL https://doi.org/10.1093/llc/fqs048.

EVERITT, B.S., LANDAU, S., LEESE, M., STAHL, D., (2011). Cluster Analysis. John Wiley & Sons, Chichester. ISBN: 978-0-470-74991-3.

GOLLEDGE, R. G., RUHTON, G., (1972). Multidimensional Scaling: Review and Geographical Applications, Technical Paper No. 10. Association of American Geographers, WASHINGTON D. C., URL http://files.eric.ed.gov/fulltext/ED110362.pdf.

GOWER, J. C., (1966). Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika, Vol. 53, Issue 3-4, pp. 325–328, URL https://doi.org/10.1093/biomet/53.3-4.325.

GRYSZEL, P., WALESIAK, M., (2014). Zastosowanie uogólnionej miary odległości GDM w ocenie atrakcyjności turystycznej powiatów Dolnego Śląska [The Application of the General Distance Measure (GDM) in the Evaluation of Lower Silesian Districts' Attractiveness], Folia Turistica, 31, pp. 127–147, URL http://www.folia-turistica.pl/attachments/article/402/FT_31_2014.pdf.

HERFINDAHL, O. C., (1950). Concentration in the Steel Industry, Doctoral thesis, Columbia University.

HIRSCHMAN, A. O., (1964). The Paternity of an Index, The American Economic Review, Vol. 54, No. 5, pp. 761-762, URL http://www.jstor.org/stable/1818582.

JAJUGA, K., WALESIAK, M., (2000). Standardisation of Data Set under Different Measurement Scales, In: Decker, R., Gaul, W., (Eds.), Classification and Information Processing at the Turn of the Millennium, 105-112. Springer-Verlag, Berlin, Heidelberg, URL http://dx.doi.org/10.1007/978-3-642-57280-7_11.

JAJUGA, K., WALESIAK, M., BĄK, A., (2003). On the General Distance Measure, in Schwaiger, M., Opitz, O., (Eds.), Exploratory Data Analysis in Empirical Research. Berlin, Heidelberg: Springer-Verlag, pp. 104–109, URL http://dx.doi.org/10.1007/978-3-642-55721-7_12.

LEMON, J., at al., (2017). plotrix: Various Plotting Functions. R package version 3.6-5, URL http://CRAN.R-project.org/package=plotrix.

MAIR, P., De LEEUW, J., BORG, I., GROENEN, P. J. F., (2017). smacof: Multidimensional Scaling. R package version 1.9-6, URL http://CRAN.R-project.org/package=smacof.

MARCUSSEN, C., (2014). Multidimensional scaling in tourism literature, Tourism Management Perspectives, Vol. 12, October, pp. 31–40, URL http://dx.doi.org/10.1016/j.tmp.2014.07.003.

PINKLEY, R.L., GELFAND, M.J., DUAN, L., (2005). When, Where and How: The Use of Multidimensional Scaling Methods in the Study of Negotiation and Social Conflict. International Negotiation, Vol. 10, Issue 1, pp 79–96, URL http://dx.doi.org/10.1163/1571806054741056.

R DEVELOPMENT CORE TEAM, (2017). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org.

STEVENS, S. S., (1946). On the Theory of Scales of Measurement. Science, Vol. 103, No. 2684, pp. 677–680, URL http://dx.doi.org/10.1126/science.103.2684.677.

TAKANE, Y., (2007). Applications of multidimensional scaling in psychometrics. In Rao, C.R., Sinharay, S. (Eds.), Handbook of Statistics, Vol. 26, Pyschometrics, Elsevier, Amsterdam, ISBN**:** 9780444521033, pp. 359–400.

TORGERSON, W. S., (1952). Multidimensional scaling: I. Theory and method. Psychometrika, Vol. 17, Issue 4, pp. 401–419, URL https://link.springer.com/article/10.1007/BF02288916.

WALESIAK, M., (2016). Visualization of Linear Ordering Results for Metric Data with the Application of Multidimensional Scaling, Ekonometria [Econometrics], 2 (52), pp. 9–21,
URL http://dx.doi.org/10.15611/ekt.2016.2.01.

WALESIAK, M., DUDEK, A., (2017a). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. R package version 0.45-2, URL http://CRAN.R-project.org/package=clusterSim.

WALESIAK, M., DUDEK, A., (2017b). mdsOpt: Searching for Optimal MDS Procedure for Metric Data. R package version 0.1-4, URL http://CRAN.R-project.org/package=mdsOpt.

ZELIAŚ, A., (2002). Some Notes on the Selection of Normalisation of Diagnostic Variables, Statistics in Transition, 5 (5), pp. 787–802.