mdsOpt – Searching for Optimal MDS Procedure for Metric and Interval-Valued Data

Marek Walesiak
Wrocław University of Economics

Andrzej Dudek
Wrocław University of Economics

Abstract

In multidimensional scaling (MDS) carried out on the basis of a metric data matrix (interval, ratio) or interval-valued data table three approaches can be distinguished: classic-to-classic – for metric data, symbolic-to-classic and symbolic-to-symbolic – for interval-valued data. The article presents the **mdsOpt** package which helps to solve the problem of choosing the optimal multidimensional scaling procedure. It uses two criteria for selecting the optimal multidimensional scaling procedure: Kruskal's *Stress*-1 fit measure (*I-Stress* in case of symbolic-to-symbolic approach) and Hirschman-Herfindahl *HHI* index calculated based on Stress per point (Interval stress per box in case of symbolic-to-symbolic approach) values. In first part three possible approaches are characterised with theoretical background of used methods and the relationships between **mdsOpt** package and existing R packages. Second part explains procedure and criteria for selection of the optimal multidimensional scaling procedure for metric and interval-valued data. The last contains in details the usage of package and examples (R scripts) on real data sets related to tourist attractiveness of Polish voivodships (provinces) and Lower-Silesian counties.

Keywords: multidimensional scaling, metric and interval-valued data, tourist attractiveness, HHI index, R.

1. Introduction

1.1. The aim of multidimensional scaling

Classical multidimensional scaling (MDS) 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 and Groenen (2005), p. 3). Classical MDS requires that each entry of dissimilarity matrix be a single numerical value. Dissimilarity between object i and object k can be fuzzy (Groenen, Winsberg, Rodriguez, and Diday (2006), p. 361). The fuzzy dissimilarity is represented by an interval and $n \times n$ dissimilarity matrix is an interval of values $\left[\delta_{ik}^{l}, \delta_{ik}^{u}\right]$, where $\delta_{ik}^{l}(\delta_{ik}^{u})$ denotes the lower (upper) bound of the dissimilarity of objects i and k in m-dimensional space. Multidimensional scaling of interval dissimilarities represents the lower and upper bounds of the dissimilarities as distances between hypercubes (rectangles in two-dimensional space and cubes in three-dimensional space). 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, and Duan (2005)), linguistics (Embleton, Uritescu, and Wheeler (2013)), marketing research (Cooper (1983)), tourism (Marcussen (2014)), musicology (McAdams, Winsberg, Donnadieu, De Soete, and Krimphoff (1995)).

1.2. The approaches in multidimensional scaling via mdsOpt package

In MDS carried out on the basis of a metric data matrix (interval, ratio) or interval-valued data table, via **mdsOpt** package, three approaches can be distinguished:

1. Classic-to-classic – for metric data:

$$\mathbf{X} = [x_{ij}]_{n\times m} \to \mathbf{Z} = [z_{ij}]_{n\times m} \to [\delta_{ik}(\mathbf{Z})]_{nxn} \to f : [\delta_{ik}(\mathbf{Z}) \to d_{ik}(\mathbf{V})] \to \mathbf{V} = [v_{ij}]_{n\times q}, \quad (1)$$

where: x_{ij} – the value of the j-th variable for the i-th object, z_{ij} – the normalized value of the j-th variable for the i-th object, $i, k = 1, \ldots, n$ – the number of the object, $j = 1, \ldots, m$ – the number of variable, $[\delta_{ik}(\mathbf{Z})]_{nxn}$ – a distance matrix (dissimilarities) between objects in m-dimensional space (distances between objects are calculated via e.g. city-block, Euclidean, Chebyshev, squared Euclidean), $[d_{ik}(\mathbf{V})]$ – a distance matrix between objects in q-dimensional space ($q < m$), f – function which mapping distances in m-dimensional space $\delta_{ik}(\mathbf{Z})$ into corresponding distances $d_{ik}(\mathbf{V})$ in q-dimensional space, $\mathbf{V} = [v_{ij}]_{n\times q}$ – data matrix in q-dimensional space.

The starting point of multidimensional scaling in classic-to-classic approach is a metric data matrix $\mathbf{X} = [x_{ij}]_{n\times m}$, for which observations are obtained from secondary data sources. It is typical situation in socio-economic research. Methods of determining the distance matrix $[\delta_{ik}]$ can be divided into direct (typically result from similarity ratings on object pairs, from rankings, or from card-sorting tasks) and indirect (see (1)) methods (see e.g. Borg and Groenen (2005), pp. 111-133).

2. Symbolic-to-classic – for interval-valued data:

$$\begin{gathered}\mathbf{X} = [x^l_{ij}, x^u_{ij}]_{n\times m} \to \mathbf{Z} = [z^l_{ij}, z^u_{ij}]_{n\times m} \to [\delta_{ik}(\mathbf{Z})]_{nxn} \to \\ f : [\delta_{ik}(\mathbf{Z}) \to d_{ik}(\mathbf{V})] \to \mathbf{V} = [v_{ij}]_{n\times q},\end{gathered} \quad (2)$$

where: $\mathbf{X} = [x^l_{ij}, x^u_{ij}]_{n\times m}$ – an interval-valued data table in m-dimensional space ($x^l_{ij} \leq x^u_{ij}$), x^l_{ij} (x^u_{ij}) – the lower (upper) bound of interval, $\mathbf{Z} = [z^l_{ij}, z^u_{ij}]_{n\times m}$ – the normalized interval-valued data table in m-dimensional space, $\delta_{ik}(\mathbf{Z})$ – a distance matrix (dissimilarities) between objects in m-dimensional space (distances between objects are calculated via distance measures for interval-valued data – see Table 3).

3. Symbolic-to-symbolic – for interval-valued data:

$$\begin{gathered}\mathbf{X} = [x^l_{ij}, x^u_{ij}]_{n\times m} \to \mathbf{Z} = [z^l_{ij}, z^u_{ij}]_{n\times m} \to \left[\delta^l_{ik}, \delta^u_{ik}\right] \to \\ f : \left(\left[\delta^l_{ik}, \delta^u_{ik}\right] \to \left[d^l_{ik}, d^u_{ik}\right]\right) \to \mathbf{V} = [v^l_{ij}, v^u_{ij}]_{n\times q},\end{gathered} \quad (3)$$

where: $\delta_{ik}^l(\delta_{ik}^u)$ – the lower (upper) bound of the dissimilarity of objects i and k in m-dimensional space, $d_{ik}^l(d_{ik}^u)$ – the lower (upper) bound of the dissimilarity of objects i and k in q-dimensional space, f – function which represent the lower and upper bounds of the dissimilarities by minimum and maximum distances between rectangles (cubes in three-dimensional space) as well as possible distances in the sense of least-squares (Groenen *et al.* (2006), p. 363), $\mathbf{V} = [v_{ij}^l, v_{ij}^u]_{n \times q}$ – an interval-valued data table in q-dimensional space.

In symbolic-to-classic and symbolic-to-symbolic approaches we assume that the starting point of multidimensional scaling is data table $\mathbf{X} = [x_{ij}^l, x_{ij}^u]_{n \times m}$ $(x_{ij}^l \leq x_{ij}^u)$. In article (Gioia and Lauro (2006), p. 344) we can find different kind of data that in real life are of interval type:

- financial data (e.g. opening value and closing value in a session),
- customer satisfaction data (expected or perceived characteristic of the quality of a product),
- tolerance limits in quality control,
- confidence intervals of estimates from sample surveys,
- query on a database.

Additional suggestions about real life interval-valued data we can find in Brito, Noirhomme-Fraiture, and Arroyo (2015):

- high–low intervals of financial prices,
- some questions in the questionnaire surveys (e.g. age, income, time spent).

Interval-valued data we can obtain by generalization of classical single-valued variables into interval-valued variables (see e.g. Bock (2000), pp. 43-44). For example, 380 Polish counties are described by 9 metric variables (see the second demonstration example). Counties are part of 16 Polish voivodships. After aggregation of data from counties to voivodships, interval-valued data are obtained. Interval of a given variable for the voivodship covers all or selected (e.g. from first to ninth decile, from first to third quartile) observations from counties.

1.3. The main idea of the mdsOpt package

The authors of the monograph (Borg, Groenen, and Mair (2013); Borg, Groenen, and Mair (2018), 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 *et al.* (2018), pp. 85-86) "The Stress value is, however, merely a technical index, a target criterion for an optimization algorithm. An MDS solution can be robust and replicable, even if its Stress value is high" and "Stress 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 (...) The least one can do is to take a look at the Stress-per-point values". Considering that we should take into account stress per point values (Borg and Mair (2017)) and Shepard diagram (Mair, Borg, and Rusch (2016); De Leeuw and Mair (2015)) for classic-to-classic and symbolic-to-classic approaches or the *I-Stress* per box index (*ispb*) and the *I-dist* diagram for symbolic-to-symbolic approach.

1.4. Criteria for selection of the optimal MDS procedure

To solve the problem of choosing the optimal multidimensional scaling procedure in:

- Classic-to-classic and symbolic-to-classic approaches two criteria were applied in **mdsOpt** package: Kruskal's *Stress*-1 (standardized residual sum of squares) fit measure and the Hirschman-Herfindahl *HHI* index, calculated based on Stress per point values (*spp*).
- Symbolic-to-symbolic approach two criteria were applied in **mdsOpt** package: *I-Stress* fit measure and the Hirschman-Herfindahl *HHI* index, calculated based on *I-Stress* per box index values (*ispb*).

1.5. Package mdsOpt versus other packages

The algorithms implemented in the **mdsOpt** package have not been used in other R program packages so far and it can be treated as a complementary package for well-known libraries **smacof** (Mair, De Leeuw, Borg, and Groenen (2019); De Leeuw and Mair (2009)) and **smds** (Terada and Groenen (2015)), extending theirs possibilities. The relationships between **mdsOpt** and other R packages present Table 1.

Additionally **mdsOpt** contain functions for calculation of *I-Stress* per box index (*ispb*) and charting *I-dist* diagram for interval-valued data.

2. Selection of the optimal multidimensional scaling procedure

The article proposes a solution that allows the optimal multidimensional scaling procedure, for metric and interval-valued data, to be chosen.

2.1. Basic decision problems

For classic-to-classic and symbolic-to-classic approaches the study uses the function smacofSym of **smacof** package of R program. In smacofSym function basic decision problems involve the following selection:

– normalization method (the analysis include 18 normalization methods – see Table 2),

– distance measure: 5 for metric data (Manhattan, Euclidean, Chebyshev, Squared Euclidean, GDM1[1] – see e.g. Everitt, Landau, Leese, and Stahl (2011), pp. 49-50) and 4 for interval-valued data (see Table 3),

– MDS model (the analysis include 3 MDS models: ratio, interval, polynomial).

For symbolic-to-symbolic approach the study uses the function IMDS of **smds** package. In function IMDS basic decision problems involve the following selection:

– normalization method – the analysis include 18 normalization methods,

[1] Cf. Jajuga, Walesiak, and Bąk (2003).

MDS approach		
Classic-to-classic	Symbolic-to-classic	Symbolic-to-symbolic
Type of data		
metric (ratio, interval)	interval-valued	interval-valued
Functions of **mdsOpt** package		
optSmacofSym_mMDS	optSmacofSymInterval	optIscalInterval
Decision problem 1: normalization method		
clusterSim (data.Normalization) **base** (R Core Team (2019)) (scale)	**clusterSim** (interval_normalization)	**clusterSim** (interval_normalization)
Decision problem 2: distance measure		
Manhattan, Euclidean, Chebyshev, Squared Euclidean, GDM1 **stats** (R Core Team (2019)) (dist) **clusterSim** (dist.GDM)	Ichino-Yaguchi, Euclidean Ichino-Yaguchi, Hausdorff, Euclidean Hausdorff **clusterSim** (dist.Symbolic)	–
Decision problem 3: MDS model / optimization method		
ratio, interval, polynomial **smacof** (smacofSym)	ratio, interval, polynomial **smacof** (smacofSym)	majorization-minimization (MM), quasi-Newton (BFGS) **smds** (IMDS)

Table 1: Relationships between **mdsOpt** and other R packages

– optimization method – the analysis include 2 methods: the majorization minimization algorithm “MM” (Groenen *et al.* (2006), p. 366); quasi-Newton method “BFGS” (Nash (1990), chapter 15).

Table 2 presents normalization methods, given by linear formula (4), which were used in the selection of the optimal MDS procedure (see Jajuga and Walesiak (2000), pp. 106-107):

$$z_{ij} = b_j x_{ij} + a_j = \frac{x_{ij} - A_j}{B_j} = \frac{1}{B_j} x_{ij} - \frac{A_j}{B_j} (b_j > 0), \tag{4}$$

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.

The normalization of variables is carried out when the variables describing the analyzed objects are metric or interval-valued. The purpose of normalization is to achieve the comparability of variables (Milligan and Cooper (1988)).

For classical metric data an observation on the j-th variable for the i-th object in a data matrix $\mathbf{X} = [x_{ij}]_{n \times m}$ is expressed as one real number. Column 1 in Table 2 presents the type

Type	Method	Parameter	
		B_j	A_j
n1	Standardization	s_j	$\overline{x}_j$
n2	Positional standardization	mad_j	med_j
n3	Unitization	r_j	$\overline{x}_j$
n3a	Positional unitization	r_j	med_j
n4	Unitization with zero minimum	r_j	$\underset{i}{min}\,\{x_{ij}\}$
n5	Normalization in range [–1; 1]	$\underset{i}{max}\,\lvert x_{ij} - \overline{x}_j\rvert$	$\overline{x}_j$
n5a	Positional normalization in range [–1; 1]	$\underset{i}{max}\,\lvert x_{ij} - med_j\rvert$	med_j
n6	Quotient transformations	s_j	0
n6a		mad_j	0
n7		r_j	0
n8		$\underset{i}{max}\,\{x_{ij}\}$	0
n9		$\overline{x}_j$	0
n9a		med_j	0
n10		$\sum_{i=1}^{n} x_{ij}$	0
n11		$\sqrt{\sum_{i=1}^{n} x_{ij}^2}$	0
n12	Normalization	$\sqrt{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}$	$\overline{x}_j$
n12a	Positional normalization	$\sqrt{\sum_{i=1}^{n} (x_{ij} - med_j)^2}$	med_j
n13	Normalization with zero being the central point	$r_j/2$	m_j

$\overline{x}_j$ – mean for the j-th variable, s_j – standard deviation for the j-th variable, r_j – range for the j-th variable, $m_j = \left(\underset{i}{max}\,\{x_{ij}\} + \underset{i}{min}\,\{x_{ij}\}\right)/2$ – 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.

Table 2: Normalization methods (based on Jajuga and Walesiak (2000); Walesiak (2018))

of normalization method adopted as the function data.Normalization of **clusterSim** package (Walesiak and Dudek (2019)). 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.

For interval-valued variables each cell x_{ij} in a data table represents the interval $x_{ij} = [x_{ij}^l, x_{ij}^u]$ ($x_{ij}^l \leq x_{ij}^u$). Interval-valued data require a special normalization approach. The lower and upper bound of the interval of the j-th variable for n objects are combined into one vector containing $2n$ observations. This approach makes it possible to apply normalization methods used for classical metric data. Other approaches to normalization of interval-valued data are presented in (Młodak (2014)). After normalization process observations on each variable from 1 to n create the lower bound of intervals while observations from $n+1$ to $2n$ create the upper bound. The data were normalized using the interval_normalization function from the **clusterSim** package.

Table 3 presents selected distance measures for interval-valued data that have been used in the selection of the optimal multidimensional scaling procedure. The methods for calculating these distances are available in dist.symbolic function of **clusterSim** package.

Symbol	Name	Distance measure $\delta_{ik}(\mathbf{Z})$
U_2_q1	Ichino-Yaguchi $q=1, \gamma=0.5$	$\sum_{j=1}^{m} \varphi(z_{ij}, z_{kj})$
U_2_q2	Euclidean Ichino-Yaguchi $q=2, \gamma=0.5$	$\sqrt{\sum_{j=1}^{m} \varphi(z_{ij}, z_{kj})^2}$
H_q1	Hausdorff $q=1$	$\sum_{j=1}^{m} \left[max \left(\left\lvert z_{ij}^{l} - z_{kj}^{l} \right\rvert, \left\lvert z_{ij}^{u} - z_{kj}^{u} \right\rvert \right) \right]$
H_q2	Euclidean Hausdorff $q=2$	$\left\{ \sum_{j=1}^{m} \left[max \left(\left\lvert z_{ij}^{l} - z_{kj}^{l} \right\rvert, \left\lvert z_{ij}^{u} - z_{kj}^{u} \right\rvert \right) \right]^2 \right\}^{1/2}$

$\varphi(z_{ij}, z_{kj}) = |z_{ij} \oplus z_{kj}| - |z_{ij} \otimes z_{kj}| + \gamma(2 \cdot |z_{ij} \otimes z_{kj}| - |z_{ij}| - |z_{kj}|)$; || – length of interval, $z_{ij} \oplus z_{kj} = z_{ij} \cup z_{kj}$, $z_{ij} \otimes z_{kj} = z_{ij} \cap z_{kj}$.

Table 3: Distance measures for interval-valued data (based on Billard and Diday (2006), pp. 244-246; Esposito *et al.* (2000), pp. 165-185; Ichino and Yaguchi (1994))

2.2. Stages in selecting the optimal procedure for MDS

The initial point of the application of smacofSym function is to determine e.g. the following values of arguments (all parameters can be changed by the user):

– initial configuration (“torgerson” classical scaling starting solution),

– convergence criterion (eps=1e-06),

– maximum number of iterations (itmax=1000).

The initial point of the application of IMDS function of **smds** package is to determine e.g. the following values of arguments (all parameters can be changed by the user):

– initial configuration (the hyper-rectangles with centres assigned as the result of classical multidimensional scaling of primary space interval centres and vertices distant from the centres by one),

– convergence criterion (eps=1e-5),

– maximum number of iterations (maxit=1000).

Selecting 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:

- In classic-to-classic approach 10 normalization methods, 5 distance measures and 4 MDS models (mspline model – polynomial function of second and third degree), there are 200 multidimensional scaling procedures.

Due to the fact that the groups of A, B, C and D (see Table 4) 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).

- In symbolic-to-classic approach 18 normalization methods, 4 distance measures for interval-valued data and 4 MDS models, there are 288 multidimensional scaling procedures.
- In symbolic-to-symbolic approach 18 normalization methods and 2 optimization methods there are 36 multidimensional scaling procedures.

Groups	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.

Table 4: The groups of normalization methods resulting in identical distance matrices (Walesiak and Dudek (2017))

3. Multidimensional scaling is performed for each procedure separately. It then orders the procedures by increasing:

- *Stress*-1 fit measure in classic-to-classic and in symbolic-to-classic approaches (see e.g. Borg *et al.* (2018), p. 32):

$$\textit{Stress}\text{-}1_p = \sqrt{\sum_{i<k}\left[d_{ik}\left(\mathbf{V}\right)-\widehat{d}_{ik}\right]^2 / \sum_{i<k} d_{ik}^2\left(\mathbf{V}\right)}, \tag{5}$$

where: p – multidimensional scaling procedure number, $\widehat{d}_{ik}$ – d-hats, disparities, target distances or pseudo distances (see Borg and Groenen (2005), p. 199), $\widehat{d}_{ik} = f(\delta_{ik})$ by defining f in different ways (ratio, interval, polynomial MDS).

- *I-Stress* fit measure in symbolic-to-symbolic approach (Groenen *et al.* (2006), p. 363):

$$\textit{I-Stress}_p = \frac{\sum_{i<k}^{n} w_{ik}\left[\delta_{ik}^u - d_{ik}^u\right]^2 + \sum_{i<k}^{n} w_{ik}\left[\delta_{ik}^l - d_{ik}^l\right]^2}{\sum_{i<k}^{n} w_{ik}\left[\delta_{ik}^u\right]^2 + \sum_{i<k}^{n} w_{ik}\left[\delta_{ik}^l\right]^2}, \tag{6}$$

where: $\delta_{ik}^l, \delta_{ik}^u$ (d_{ik}^l, d_{ik}^u) – the lower and upper bound of the dissimilarity in m-dimensional space (q-dimensional space), w_{ik} – nonnegative weight (in general $w_{ik} = 1$).

4. Based on Stress per point (*spp*) values (Stress contribution in percentages), the Hirschman-Herfindahl index is calculated (Herfindahl (1950); Hirschman (1964)) in classic-to-classic and in symbolic-to-classic approaches:

$$HHI_p = \sum_{i=1}^{n} spp_{pi}^2, \tag{7}$$

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

Based on Interval stress per box (*ispb*) values (Interval Stress contribution in percentages), the Hirschman-Herfindahl index is calculated in symbolic-to-symbolic approach:

$$HHI_p = \sum_{i=1}^{n} ispb_{pi}^2, \tag{8}$$

where:

$$ispb_i = \frac{\left(\sum_{k=1}^{n} w_{ik} \left[\delta_{ik}^u - d_{ik}^u\right]^2 + \sum_{k=1}^{n} w_{ik} \left[\delta_{ik}^l - d_{ik}^l\right]^2\right)/n}{\sum_{i=1}^{n} \left[\left(\sum_{k=1}^{n} w_{ik} \left[\delta_{ik}^u - d_{ik}^u\right]^2 + \sum_{k=1}^{n} w_{ik} \left[\delta_{ik}^l - d_{ik}^l\right]^2\right)/n\right]} \cdot 100$$

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. Maximal value appears when summary fit measure (*Stress*-1, *I-Stress*) 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.

5. The chart with *Stress*-1_p (*I-Stress*$_p$) fit measure value on x-axis and HHI_p index on y-axis for p procedures of multidimensional scaling is drawn.

6. The maximal acceptable value of *Stress*-1 (*I-Stress*) is assumed as *cs* (may be calculated as a midrange or median of *Stress*-1 (*I-Stress*)). For all multidimensional scaling procedures, for which *Stress*-$1_p \leq cs$ (*I-Stress*$_p \leq cs$), we choose the one for each occurs $\underset{p}{min}\{HHI_p\}$.

7. 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 (*I-dist* diagram) and *Stress* (*I-Stress*) plot, 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).

3. Using the package – examples

3.1. Metric data (classic-to-classic approach)

In first example we will find the optimal solution for classic-to-classic MDS approach. The package **mdsOpt** contains dataset called data_lower_silesian referring to the attractiveness level of 31 objects (29 Lower Silesian counties, Pattern and Anti-pattern object) described by 16 metric variables:

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

x2 – number of nights spent daily by resident tourists 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 zloty amounts (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 the county in the years 2011-2012, per 1 km^2 of the county area.

The statistical data were collected in 2012 and come from the Local Data Bank of the Statistics Poland (GUS); the data for x7 variable only were obtained from the regional conservation officer. Variables x1-x3, x7, x8, x10-x16 represent stimulants (where higher values are more p rred), variables x4, x5 and x6 take the form of destimulants (where lower values are more preferred), and x9 is a nominant (nominal value is the best value and lies within range of variable – 50% level was adopted as the optimal one). Variable x9 was transformed into a stimulant. The coordinates of a Pattern object cover the most preferred preference variable values (maximum for stimulant, minimum for destimulant). The coordinates of an Anti-pattern object cover the least preferred preference variable values (minimum for stimulant, maximum for destimulant).

First we load package and dataset.

```
R> library(mdsOpt)
R> data(data_lower_silesian)
```

Then set the normalizations methods, distance measures and MDS models used in selection of optimal MDS procedure.

```
R> metnor<-c("n1","n2","n3","n5","n5a","n8","n9","n9a","n11","n12a")
R> metscale<-c("ratio","interval","mspline")
R> metdist<-c("euclidean","manhattan","seuclidean","maximum","GDM1")
```

The normalizations methods, distance measures and MDS models are used in next step (please notice that model mspline is used twice with spline.degree parameter equals 2 or 3) for selecting the optimal multidimensional scaling procedure.

```
R> res<-optSmacofSym_mMDS(data_lower_silesian,normalizations=metnor,
+ distances=metdist,mdsmodels=metscale,spline.degrees=c(2:3),outDec=".",
+ stressDigits=6,HHIDigits=2)
```

The results contain 200 rows (10 normalization methods x 5 distance measures x 4 MDS models) each describing one procedure with the six columns: Normalization method, MDS model, Spline degree, Distance measure, STRESS 1, HHI spp. The values are ordered by STRESS 1 value.

Before displaying the result we need to change the max.print system option to value greater or equals 1200 (10 normalization methods x 5 distance measures x 4 MDS models x 6 columns).

```
R> options(max.print=1200)
R> print(res)

       Normalization MDS model  Spline  Distance     STRESS 1   HHI spp
       method                   degree  measure
  [1,] "n9a"         "mspline"  "3"     "euclidean"  "0.026339" " 821.90"
  [2,] "n9a"         "mspline"  "2"     "euclidean"  "0.026451" " 856.47"
  [3,] "n9a"         "mspline"  "2"     "seuclidean" "0.026967" " 791.68"
  ...
[198,] "n8"          "ratio"    ""      "maximum"    "0.261772" " 414.10"
[199,] "n3"          "ratio"    ""      "maximum"    "0.265246" " 414.13"
[200,] "n5"          "ratio"    ""      "maximum"    "0.266663" " 404.71"
```

Then we convert *Stress*-1 and *HHI* values to numeric vectors.

```
R> stress<-as.numeric(res[,"STRESS 1"])
R> hhi<-as.numeric(res[,"HHI spp"])
```

The maximal acceptable *cs* value is calculated as a mid-range of *Stress*-1 values.

```
R> cs<-(min(stress)+max(stress))/2
R> print(cs)
[1] 0.146501
```

Then the best MDS procedure from all combinations is chosen.

```
# Elements of optimal MDS procedure
R> t<-findOptimalSmacofSym(res,cs)
R> print(t)

\$`Nr`
[1] 117
\$Normalization_method
[1] "n12a"
\$MDS_model
[1] "interval"
\$Spline_degree
[1] ""
\$Distance_measure
[1] "euclidean"
\$STRESS_1
[1] 0.132176
\$HHI_spp
[1] 420.74
```

In next step we can plot dependency between *Stress*-1 and *HHI* index (see Figure 1) with best solution marked by red circle and finally we choose the MDS solution that satisfies condition *Stress*-1≤*cs* and minimizes *HHI*.

```
# Plot dependency between Stress-1 and HHI index
R> plot(stress[-t$Nr],hhi[-t$Nr],xlab="Stress-1",ylab="HHI",
+ type="n",font.lab=3)
R> text(stress[-t$Nr],hhi[-t$Nr],labels=(1:nrow(res))[-t$Nr])
R> abline(v=cs,col="red")
R> points(stress[t$Nr],hhi[t$Nr],cex=5,col="red")
R> text(stress[t$Nr],hhi[t$Nr],labels=(1:nrow(res))[t$Nr],col="red")
```

The results of optimal multidimensional scaling procedure (117), via below script, for 31 objects (29 Lower Silesian counties, Pattern and Anti-pattern object) according to the level of tourist attractiveness are presented on Figure 2.

```
R> library(mdsOpt)
R> data(data_lower_silesian)
R> z<-data.Normalization(data_lower_silesian,type="n12a")
R> d<-dist(z,method="euclidean")
R> res<-smacofSym(delta=d,ndim=2,type="interval")
R> par(mfrow=c(2,2),pty="s")
# Shepard Diagram
R> plot(res,plot.type="Shepard",cex.main=0.8,cex.lab=0.8,cex.axis=0.8,cex=0.2)
# Stress plot
R> spp<-sort(res$spp,decreasing=TRUE)
R> names(spp)<-order(res$spp,decreasing=TRUE)
```

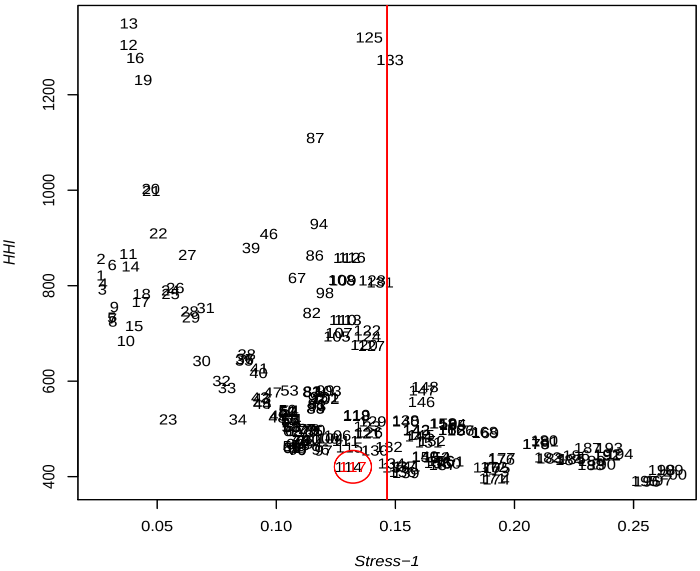

Figure 1: The values of *Stress*-1 fit measure and *HHI* index
for p multidimensional scaling procedures (with best solution marked by red circle)

```
R> plot(spp,main="Stress plot",ylab="Stress contribution in percents",
+ xlab="Objects",ylim=c(-2,30),cex=0.4,cex.main=0.8,cex.lab=0.8,cex.axis=0.8)
R> text(spp,pos=3,names(spp),cex=0.4)
# Configuration plot with bubble
R> bubsize=res$spp/length(spp)*4
R> plot(res$conf,main="Configuration plot with bubble",xlab="Dimension 1",
+ ylab="Dimension 2",cex=bubsize,cex.main=0.8,cex.lab=0.8,cex.axis=0.8,asp=1)
R> text(res$conf[,1],res$conf[,2],pos=3,1:nrow(res$conf),cex=0.7)
R> arrows(res$conf[nrow(z),1],res$conf[nrow(z),2],res$conf[nrow(z)-1,1],
+ res$conf[nrow(z)-1,2],length=0.05,col="black")
R> plot.new()
R> legend("center",paste(1:nrow(res$conf),rownames(res$conf)),
+ bty="n",cex=0.7,ncol=2,title="Legend")
```

Figure 2 (Configuration plot with bubble) presents additional the quota of each object in total error is shown by the size of radius of the circle around each object. Shepard Diagram and Stress plot confirm the correctness of the chosen scaling model. On Figure 2 (Configuration plot with bubble), the axis of the set, which means the shortest connection between Pattern and Anti-pattern object, is designated. It indicates the level of development of the tourist attractiveness of counties. Objects that are closer to Pattern object have higher level of tourist attractiveness.

To opposite to the best MDS procedure (117) we show the results for the one of the worst procedures (13): n9a normalization method, mspline of third degree MDS model, maximum (Chebyshev) distance. In relation to the previous script, changes in lines 3-5 and in Shepard

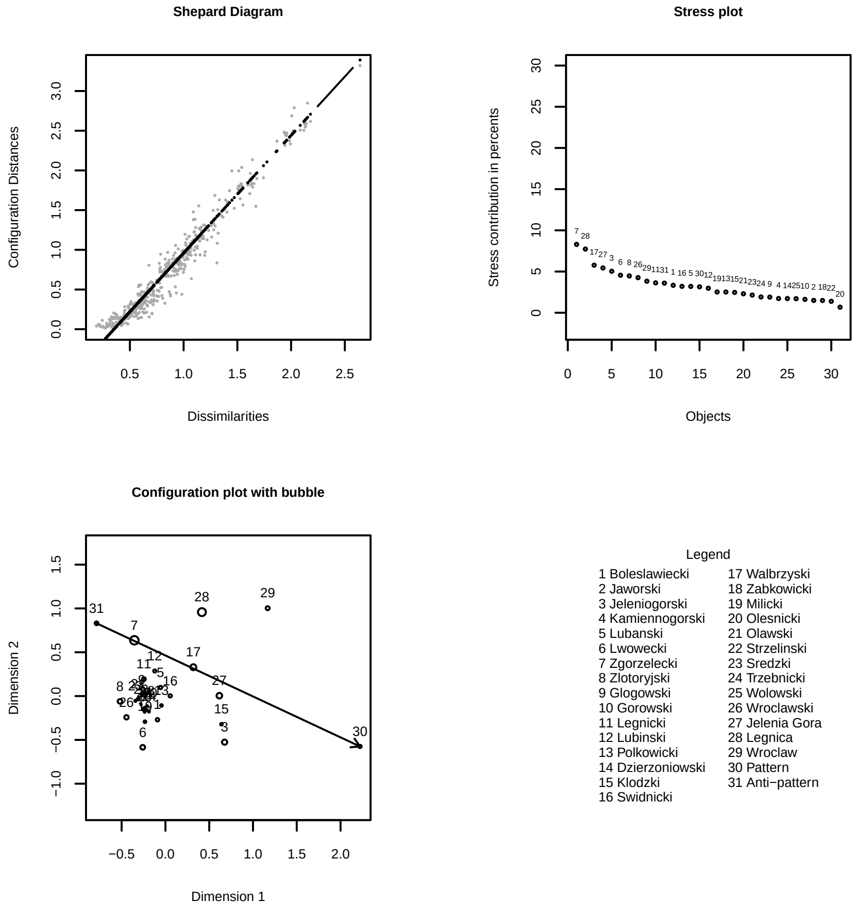

Figure 2: The results of multidimensional scaling (procedure 117) of 31 objects (29 Lower Silesian counties, Pattern and Anti-pattern) according to the level of tourist attractiveness

diagram are required.

```
R> z<-data.Normalization(data_lower_silesian,type="n9a")
R> d<-dist(z,method="maximum")
R> res<-smacofSym(delta=d,ndim=2,type="mspline",spline.degree=3)
...
# Shepard Diagram
R> plot(res,plot.type="Shepard",cex.main=0.8,cex.lab=0.8,cex.axis=0.8,cex=0.2)
R> t1<-as.matrix(res$delta)
R> t2<-as.matrix(res$confdist)
R> text(t1[7,3],t2[7,3],pos=4,"7,3",cex=0.6)
R> text(t1[31,3],t2[31,3],pos=1,"31,3",cex=0.6)
```

The results of multidimensional scaling for procedure 13 according to the level of tourist attractiveness are presented on Figure 3.

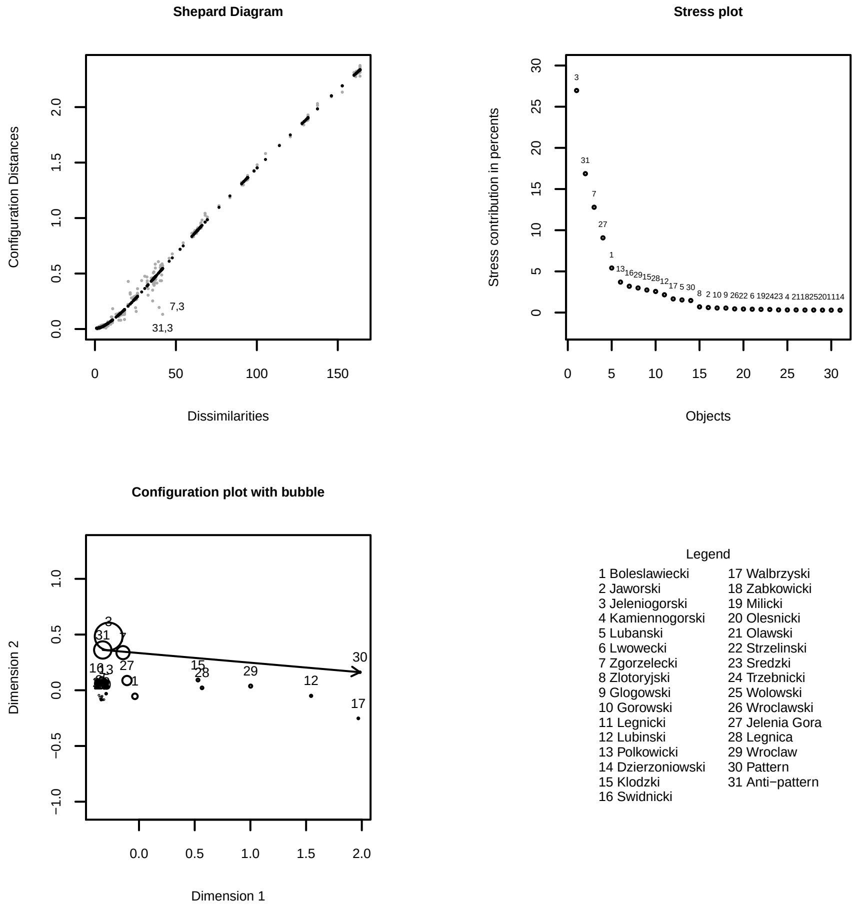

Figure 3: The results of multidimensional scaling (procedure 13) of 31 objects (29 Lower Silesian counties, Pattern and Anti-pattern) according to the level of tourist attractiveness

Overall Stress for procedure 13 (0.0381) is much better than for procedure 117 (0.1322). Figure 3 (Stress plot) exhibits that objects Jeleniogorski (3), Anti-pattern (31) and Zgorzelecki (7) contribute most to the overall Stress (56.62%). It also shows (see Shepard Diagram – in the lower left-hand corner) that two points (distance between Jeleniogorski county (3) and Anti-pattern object (31); Jeleniogorski county (3) and Zgorzelecki (7) county) are outliers. These outliers contribute over-proportionally to the total Stress. MDS configuration (Figure 3 – Configuration plot with bubble) does not represent all proximities equally good. Jeleniogorski county (3) is one of the best of Lower Silesian counties according to the level of tourist attractiveness. In 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 representation real relationships between objects.

3.2. Interval-valued data (symbolic-to-symbolic approach)

In second example we will find the optimal solution for symbolic-to-symbolic MDS approach.

The dataset data_symbolic_interval_polish_voivodships comes from **clusterSim** package. For the evaluation of tourist attractiveness of Polish voivodships (provinces) in the year 2016 a two-stage data collection has been carried out.

Step 1. Data on tourist attractiveness were collected for 380 Polish counties for the following metric variables:

x1 – beds in hotels per 1,000 inhabitants of a county,

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

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

x4 – dust pollution emission in tons per 10 km^2 of a county area,

x5 – gas pollution emission in tons per 1 km^2 of a county area,

x6 – number of criminal offences, crimes against life and health and property crimes per 1,000 inhabitants of a county,

x7 – forest cover of the county in %,

x8 – participants of mass events per 1,000 inhabitants of a county,

x9 – number of tourist economy entities (sections: I, N79) registered in the system REGON per 1,000 inhabitants of a county.

Three variables x4, x5 i x6 can be treated as destimulants. All other variables are stimulants.

Step 2. Data table has been aggregated up to the voivodships with interval-valued data as an a result. The lower bound of the interval for each variable was obtained by calculating the first quartile based on data from the counties. In turn, the upper bound of the interval was obtained by calculating the third quartile. In result dataset contains data about 18 objects (16 voivodships, Pattern and Anti-pattern) described by 9 interval-valued variables.

First we load package **mdsOpt** and dataset (please notice that there is no need to load **clusterSim** package – it is auto-loaded automatically by **mdsOpt**).

```
R> library(mdsOpt)
R> data("data_symbolic_interval_polish_voivodships")
R> data<-data_symbolic_interval_polish_voivodships
```

Then set the normalizations methods and optimization methods used in selection of optimal MDS procedure.

```
R> metnor<-c("n1","n2","n3","n3a","n4","n5","n5a","n6","n6a","n7","n8","n9",
+ "n9a","n10","n11","n12","n12a","n13")
R> methods<-c("MM","BFGS")
```

In next step we run *I-scal* algorithm for all combinations of normalization methods and optimization methods with default parameters.

```
R> res<-optIscalInterval(x=data,dataType="simple",normalizations=metnor,
+ optMethods=methods,outDec=".",stressDigits=6,HHIDigits=2)

initial value 568.744280
iter 100 stress = 33.568175
```

```
....
final (iter 1000) stress = 9.392537
stopped after 1000 iterations
initial  value 217.126687
final  value 3.943757
converged

R> print(res)

      Normalization method Opt method I-STRESS   HHI spb
 [1,] "n9a"                "MM"       "0.000087" " 746.31"
 [2,] "n9a"                "BFGS"     "0.000108" "1156.36"
 [3,] "n2"                 "BFGS"     "0.000200" " 863.13"
 ...
[34,] "n4"                 "MM"       "0.007690" "1316.30"
[35,] "n12"                "MM"       "0.008430" "1148.12"
[36,] "n12a"               "MM"       "0.009668" "1058.55"
```

The values are ordered by *I-Stress* value. Then we convert *I-Stress* and *HHI* values to numeric vectors.

```
R> Istress<-as.numeric(res[,"I-STRESS"])
R> hhi<-as.numeric(res[,"HHI spb"])
```

The maximal acceptable *cs* value is calculated as an median of *I-Stress* values.

```
R> cs<-median(Istress)
R> print(cs)

[1] 0.003215
```

Then the best MDS procedure from all combinations is chosen.

```
# Elements of optimal MDS procedure
R> t<-findOptimalIscalInterval(res,cs)
R> print(t)

$Nr
[1] 5
$Normalization_method
[1] "n2"
$Opt_method
[1] "MM"
$I_STRESS
[1] 0.000268
$HHI_spb
[1] 743.61
```

In next step we can plot dependency between *I-Stress* and *HHI* index (see Figure 4) with best solution marked by red circle and finally we choose the MDS solution that satisfies condition $I\text{-}Stress \leq cs$ and minimizes *HHI*.

```
# Plot dependency between I-Stress and HHI index
R> plot(Istress[-t$Nr],hhi[-t$Nr], xlab="I-Stress",ylab="HHI",
+ type="n",font.lab=3)
R> text(Istress[-t$Nr],hhi[-t$Nr],labels=(1:nrow(res))[-t$Nr])
R> abline(v=cs,col="red")
R> points(Istress[t$Nr],hhi[t$Nr], cex=5,col="red")
R> text(Istress[t$Nr],hhi[t$Nr],labels=(1:nrow(res))[t$Nr],col="red")
```

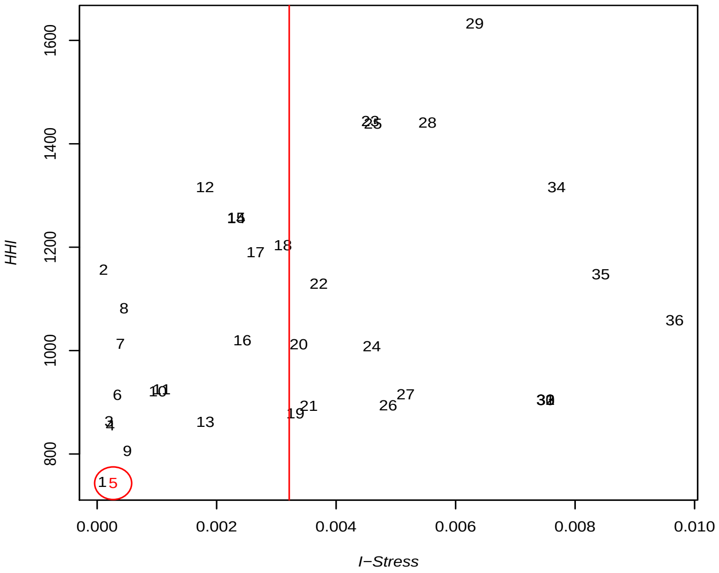

Figure 4: The values of *I-Stress* fit measure and *HHI* index
for *p* multidimensional scaling procedures (with best solution marked by red circle)

Now we will display the results of the best MDS procedure (5). First we need to load **smds** library.

```
R> library(smds)
```

The results of optimal multidimensional scaling procedure (5), via below script, for 18 objects (16 voivodships, Pattern and Anti-pattern object) according to the level of tourist attractiveness are presented on Figure 5.

```
R> library(mdsOpt)
R> data("data_symbolic_interval_polish_voivodships")
R> data<-data_symbolic_interval_polish_voivodships
```

```
R> normalized<-interval_normalization(x=data,dataType="simple",type="n2")
R> x<-normalized$simple[,,1];y<-normalized$simple[,,2]
R> my.idiss<-idistBox(X=(x+y)/2,R=(y-x)/2)
R> cmat<-(my.idiss[2,,]+my.idiss[1,,])/2
R> iniX<-cmdscale(as.dist(cmat),k=2)
R> n=dim(my.idiss)[2]
R> iniR<-matrix(rep(1,n*2),nrow=n,ncol=2)
R> res.box<-IMDS(IDM=my.idiss,p=2,model="box",opt.method="MM",
+ report=1001,ini=list(iniX,iniR))
R> x_l<-res.box$EIDM[1,,];x_u<-res.box$IDM[2,,]
R> y_l<-res.box$IDM[1,,];y_u<-res.box$EIDM[2,,]
R> spb<-ispb(res.box$EIDM,my.idiss)
R> HHI<-sum(spb^2)
R> par(mfrow=c(2,2),pty="s")
# I-dist diagram
R> plot(x_u,y_u, main="I-dist diagram",
+ ylab="The lower (red) and upper (green)\n configuration distances",
+ xlab="The lower (red) and upper\n (green) dissimilarities",
+ col="green",cex.main=0.8,cex.lab=0.8,cex.axis=0.8,cex=0.5)
R> points(x_l,y_l,col="red",cex=0.5)
# I-Stress plot
R> w<-sort(spb,decreasing=TRUE)
R> names(w)<-order(spb,decreasing=TRUE)
R> plot(w,main="I-Stress plot",xlab="Object",ylab="ispb in percents",
+ ylim=c(-2,25),cex=0.4,cex.main=0.8,cex.lab=0.8,cex.axis=0.8)
R> text(w,pos=3,names(w),cex=0.6)
# Configuration plot
R> x<-(res.box$X-res.box$R);y<-(res.box$X+res.box$R)
R> plot(NULL,xlim=c(min(x[,1]),max(y[,1])),ylim=c(min(x[,2]),max(y[,2])),
+ pch=1,cex=0.4,main="Configuration plot",xlab="Dimension 1",
+ ylab="Dimension 2",cex.main=0.8,cex.lab=0.8,asp=1,cex.axis=0.8)
R> rect(x[,1],x[,2],y[,1],y[,2])
R> text(res.box$X[,1],res.box$X[,2],labels=1:18,cex=0.8)
R> plot.new()
R> legend("center",legend=paste(1:dim(data)[[1]],attr(data,"row.names")),
+ bty="n",ncol=2,cex=0.65,title="Legend")
```

Figure 5 (*I-dist* diagram and *I-Stress* plot) confirms the correctness of the MDS results (Configuration plot). Objects that are closer to pattern of development have higher level of tourist attractiveness.

To opposite to the best MDS procedure (5) we show, via below script, the results for the one of the worst procedures (12) according to *HHI* index. In relation to the previous script, changes in lines 5, 13-14 and *I*-dist diagram are required.

```
R> normalized<-interval_normalization(x=data,dataType="simple",type="n5a")
...
R> res.box<-IMDS(IDM=my.idiss,p=2,model="box",opt.method="BFGS",
```

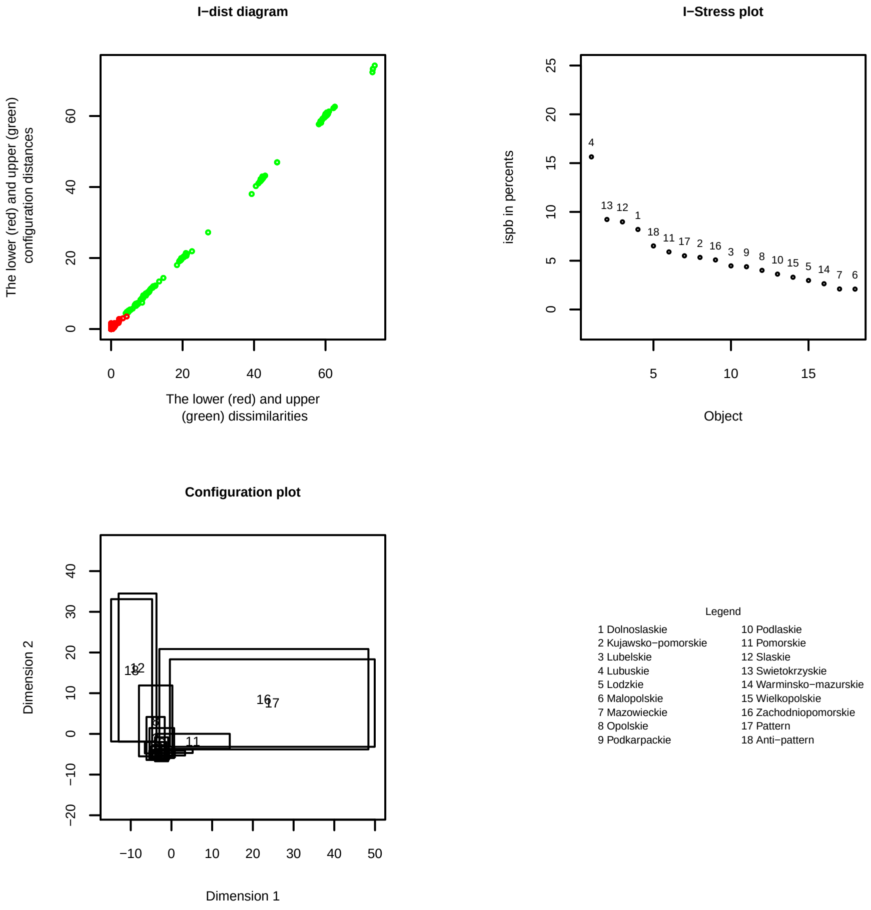

Figure 5: The results of multidimensional scaling (procedure 5) of 18 objects (16 voivodships, Pattern and Anti-pattern) according to the level of tourist attractiveness

```
+ report=1001,ini=list(iniX,iniR))
...
# I-dist diagram
R> plot(x_u,y_u, main="I-dist diagram",
+ ylab="The lower (red) and upper (green)\n configuration distances",
+ xlab="The lower (red) and upper\n (green) dissimilarities",col="green",
+ cex.main=0.8,cex.lab=0.8,cex.axis=0.8,cex=0.5)
R> points(x_l,y_l,col="red",cex=0.5)
R> text(x_u[17,16],y_u[17,16],pos=2,"17,16",cex=0.6)
R> text(x_u[17,4],y_u[17,4],pos=1,"17,4",cex=0.6)
R> text(x_l[16,9],y_l[16,9],pos=3,"16,9",cex=0.6)
```

The results of multidimensional scaling for procedure 12 according to the level of tourist attractiveness are presented on Figure 6.

Figure 6 (*I-Stress* plot) exhibits that objects Lubuskie (4), Pattern (17) and Zachodniopomorskie (16) contribute most to the overall *I-Stress* (57,68%). It also shows (see Figure

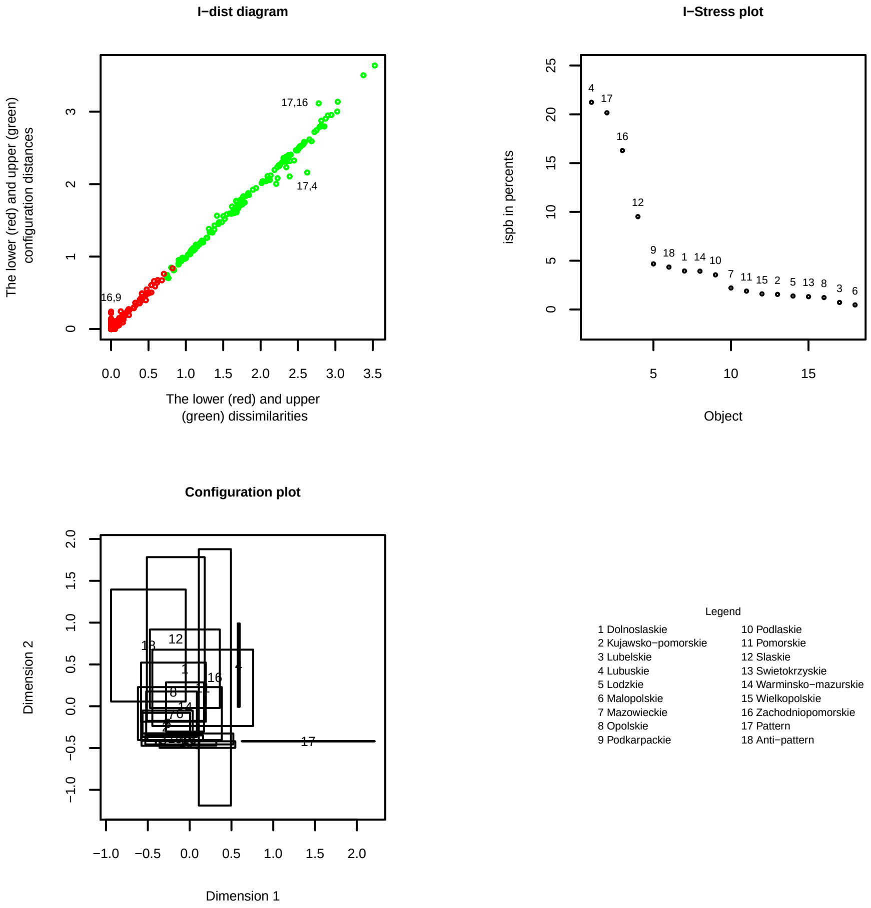

Figure 6: The results of multidimensional scaling (procedure 5) of 18 objects (16 voivodships, Pattern and Anti-pattern) according to the level of tourist attractiveness

6 – *I-dist* diagram) that some points (upper distances between Zachodniopomorskie (16) voivodship and Pattern object (17); Pattern object (17) and Lubuskie voivodship (4); lower distance between Zachodniopomorskie (16) voivodship and Podkarpackie voivodship (9)) are outliers. These outliers contribute over-proportionally to the total *I-Stress*. MDS configuration (Figure 6 – Configuration plot) does not represent all proximities equally good. Zachodniopomorskie (16) is the best of Polish voivodships according to the level of tourist attractiveness. In Figure 6 (Configuration plot) this voivodship lies further from Pattern object than Lubuskie (4). The greater the value of the HHI_p index, the worse is the effect of multidimensional scaling in terms of representation real relationships between objects.

4. Summary

The article proposes a methodology that allows the selection of the optimal MDS procedure for classical metric and interval-valued data. For classic-to-classic approach we choose best MDS procedure due to the used methods of normalization, distance measures and scaling models carried out on the basis of the metric data matrix. On the basis of this methodology research results are illustrated by first example to find the optimal procedure for multidimensional scaling of set of objects representing 29 counties in Lower Silesia according to the level of tourist attractiveness.

For symbolic-to-classic approach we choose the best MDS procedure due to the used methods of normalization, distance measures for interval-valued data and scaling models carried out on the basis of the interval-valued data table.

For symbolic-to-symbolic approach we choose the best MDS procedure due to the used methods of normalization and optimization methods carried out on the basis of the interval-valued data table. On the basis of this methodology research results are illustrated by second example to find the optimal procedure for multidimensional scaling of set of objects representing 16 Polish voivodships according to the level of tourist attractiveness.

To solve the problem of choosing the optimal multidimensional scaling procedure two criteria were applied in **mdsOpt** package Kruskal's *Stress*-1 fit measure and the Hirschman-Herfindahl *HHI* index (in classic-to-classic and symbolic-to-classic approaches) and *I-Stress* fit measure and the Hirschman-Herfindahl *HHI* index (in symbolic-to-symbolic approach).

In step 6 the maximal acceptable value of fit measures *Stress*-1 and *I-Stress* has been arbitrary assumed. It is not determined how much error distribution for each object may deviate from the uniform distribution. Among the procedures of multidimensional scaling for which *Stress*-1$\leq cs$ (*I-Stress*$\leq cs$) the one for which occurs $\underset{p}{min}\{HHI_p\}$ is selected. This constraint does not essentially limit the presented proposal, as additional criteria for acceptability such as Shepard diagram (De Leeuw and Mair (2015)) and Stress plot or *I-dist* diagram and *I-Stress* plot confirm the correctness of the MDS results.

References

Billard L, Diday E (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons, Chichester. ISBN 978-0-470-09016-9.

Bock HH (2000). "Symbolic data." In HH Bock, E Diday (eds.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, pp. 39–53. Springer-Verlag, Berlin, Heidelberg. ISBN 978-3-540-66619-6. doi:10.1007/978-3-642-57155-8.

Borg I, Groenen PJF (2005). *Modern Multidimensional Scaling. Theory and Applications*. 2nd edition edition. Springer Science+Business Media, New York. ISBN 978-0387-25150-9. doi:10.1007/0-387-28981-X.

Borg I, Groenen PJF, Mair P (2013). *Applied Multidimensional Scaling*. Springer, Heidelberg, New York, Dordrecht, London. ISBN 978-3-642-31847-4. doi:10.1007/978-3-642-31848-1.

Borg I, Groenen PJF, Mair P (2018). *Applied Multidimensional Scaling and Unfolding*. Springer, Heidelberg, New York, Dordrecht, London. ISBN 978-3-319-73470-5. doi:10.1007/978-3-319-73471-2.

Borg I, Mair P (2017). "The Choice of Initial Configurations in Multidimensional Scaling: Local Minima, Fit, and Interpretability." *Austrian Journal of Statistics*, **46**(2), 19–32. doi:10.17713/ajs.v46i2.561.

Brito P, Noirhomme-Fraiture M, Arroyo J (2015). "Editorial for Special Issue on Symbolic Data Analysis." *Advanced in Data Analysis and Classification*, **9**(1), 1–4. doi:10.1007/s11634-015-0202-1.

Cooper LG (1983). "A Review of Multidimensional Scaling in Marketing Research." *Applied Psychological Measurement*, **7**(4), 427–450. doi:10.1177/014662168300700404.

De Leeuw J, Mair P (2009). "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software*, **31**(3), 1–30. doi:10.18637/jss.v031.i03.

De Leeuw J, Mair P (2015). *Shepard Diagram.* Wiley StatsRef: Statistics Reference Online. doi:10.1002/9781118445112.stat06268.pub2.

Embleton S, Uritescu D, Wheeler ES (2013). "Defining Dialect Regions with Interpretations: Advancing the Multidimensional Scaling Approach." *Literary and Linguistic Computing*, **28**(1), 13–22. doi:10.1093/llc/fqs048.

Esposito F, Malerba D, Tamma V (2000). "Dissimilarity Measures for Symbolic Objects." In HH Bock, E Diday (eds.), *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, pp. 165–185. Springer-Verlag, Berlin, Heidelberg. ISBN 978-3-540-66619-6. doi:10.1007/978-3-642-57155-8.

Everitt B, Landau S, Leese M, Stahl D (2011). *Cluster Analysis.* John Wiley & Sons, Chichester. ISBN 978-0-470-74991-3. doi:10.1002/9780470977811.

Gioia F, Lauro CN (2006). "Principal Component Analysis on Interval Data." *Computational Statistics*, **21**(2), 343–363. doi:10.1007/s00180-006-0267-6.

Groenen PJF, Winsberg S, Rodriguez O, Diday E (2006). "I-Scal: Multidimensional Scaling of Interval Dissimilarities." *Computational Statistics & Data Analysis*, **51**(1), 360–378. doi:10.1016/j.csda.2006.04.003.

Herfindahl OC (1950). *Concentration in the Steel Industry.* Ph.D. thesis, Columbia University.

Hirschman AO (1964). "The Paternity of an Index." *The American Economic Review*, **54**(5), 761–762. URL http://www.jstor.org/stable/1818582.

Ichino M, Yaguchi H (1994). "Generalized Minkowski Metrics for Mixed Feature-type Data Analysis." *IEEE Transactions on Systems, Man, and Cybernetics*, **24**(4), 698–708. doi:10.1109/21.286391.

Jajuga K, Walesiak M (2000). "Standardisation of Data Set under Different Measurement Scales." In R Decker, W Gaul (eds.), *Classification and Information Processing at the Turn of the Millennium*, pp. 105–112. Springer-Verlag, Berlin, Heidelberg. doi:10.1007/978-3-642-57280-7_11.

Jajuga K, Walesiak M, Bąk A (2003). "On the General Distance Measure." In M Schwaiger, O Opitz (eds.), *Exploratory Data Analysis in Empirical Research*, pp. 104–109. Springer-Verlag, Berlin, Heidelberg. doi:10.1007/978-3-642-55721-7_12.

Mair P, Borg I, Rusch T (2016). "Goodness-of-fit Assessment in Multidimensional Scaling and Unfolding." *Multivariate Behavioral Research*, **51**(6), 772–789. doi:10.1080/00273171.2016.1235966.

Mair P, De Leeuw J, Borg I, Groenen PJF (2019). *smacof: Multidimensional Scaling*, R package version 2.0-0 edition. URL https://CRAN.R-project.org/package=smacof.

Marcussen C (2014). "Multidimensional Scaling in Tourism Literature." *Tourism Management Perspectives*, **12**(October), 31–40. doi:10.1016/j.tmp.2014.07.003.

McAdams S, Winsberg S, Donnadieu S, De Soete G, Krimphoff J (1995). "Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities, and Latent Subject Classes." *Psychological Res.*, **58**(3), 177–192.

Milligan GW, Cooper MC (1988). "A Study of Standardization of Variables in Cluster Analysis." *Journal of Classification*, **5**(2), 181–204.

Młodak A (2014). "On the Construction of an Aggregated Measure of the Development of Interval Data." *Computational Statistics*, **29**(5), 895–929. doi:10.1007/s00180-013-0469-7.

Nash JC (1990). *Compact Numerical Methods for Computers. Linear Algebra and Function Minimisation*. Adam Hilger, Bristol and New York. ISBN 0-85274-318-1.

Pinkley RL, Gelfand MJ, Duan L (2005). "When, Where and How: The Use of Multidimensional Scaling Methods in the Study of Negotiation and Social Conflict." *International Negotiation*, **10**(1), 79–96. doi:10.1163/1571806054741056.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Takane Y (2007). "Applications of Multidimensional Scaling in Psychometrics." In CR Rao, S Sinharay (eds.), *Handbook of Statistics (Vol. 26): Pyschometrics*, pp. 359–400. Elsevier, Amsterdam. ISBN 9780444521033.

Terada Y, Groenen PJF (2015). *smds: Symbolic Multidimensional Scaling*, R package version 1.0 edition. URL https://CRAN.R-project.org/package=smds.

Walesiak M (2018). "The Choice of Normalization Method and Rankings of the Set of Objects Based on Composite Indicator Values." *Statistics in Transition – new series*, **19**(4), 693–710.

Walesiak M, Dudek A (2017). "Selecting the Optimal Multidimensional Scaling Procedure for Metric Data with R Environment." *Statistics in Transition – new series*, **18**(3), 521–540.

Walesiak M, Dudek A (2019). *clusterSim: Searching for Optimal Clustering Procedure for a Data Set*, R package version 0.47-4 edition. URL https://CRAN.R-project.org/package=clusterSim.

Affiliation:

Marek Walesiak
Wroclaw University of Economics
Department of Econometrics and Computer Science
Komandorska 118/120
53-345 Wrocław, Poland
E-mail: marek.walesiak@ue.wroc.pl

Andrzej Dudek
Wroclaw University of Economics
Department of Econometrics and Computer Science
Komandorska 118/120
53-345 Wrocław, Poland
E-mail: andrzej.dudek@ue.wroc.pl