A comparative analysis of rankings of Polish provinces in terms of social cohesion for metric and interval-valued data

Marek Walesiak[1], Grażyna Dehnel[2]

Abstract

The article describes a comparative analysis of rankings of Polish provinces in terms of social cohesion based on metric and interval-valued data between 1st and 3rd quartiles (50% of observations), 1st and 9th deciles (80% of observations) and the minimum and maximum (100% observations). The rankings were obtained using a hybrid approach combining the use of multidimensional scaling (MDS) with linear ordering. Interval-valued variables characterise the objects of interests more accurately than metric data. Metric data are of an atomic nature, i.e. an observation of each variable is expressed as a single real number. In contrast, an observation of each interval-valued variable is expressed as an interval. Interval-valued data were derived by aggregating metric data on social cohesion at the level of districts to the province level. All observations were included in the aggregation and then outliers were omitted.

Keywords: *social cohesion, interval-valued data, multidimensional scaling, composite indicators, outliers*
JEL Classification: *C38, C43, C63*

1. Introduction and motivation

Social cohesion is a multi-faceted phenomenon. When one analyses the conceptualisation of social cohesion, one can note a clear direction of changes, reflecting the growing importance attached to the socio-cultural and political indicators, accompanied by the declining role of the economic dimension (Chan and Chan, 2006). Dickes et al. (2010) and Dickes and Valentova (2013) indicate four dimensions of social cohesion: institutional trust, solidarity, socio-cultural participation and political participation.

The article presents a comparative analysis of the rankings of Polish provinces in terms of social cohesion for metric data and three types of interval-valued data. The rankings were obtained by using a hybrid approach, which combines MDS and linear ordering. Two criteria are proposed as the basis for comparing the rankings. The first one involves cluster analysis, which is used to identify similarities and differences in the ordering of provinces in terms of social cohesion. The second one is based on the analysis of the degree to which different rankings of objects with respect to specific variables correspond to those obtained by using the aggregate measure for 4 datasets (one containing metric data and three with interval-valued data). These two approaches were then used to select a ranking that best represents the level of social cohesion in the provinces of Poland.

[1] Corresponding author: Wrocław University of Economics, Department of Econometrics and Computer Science, 3 Nowowiejska St., 58-500 Jelenia Góra, Poland, marek.walesiak@ue.wroc.pl. ORCiD 0000-0003-0922-2323.
[2] Poznan University of Economics and Business, Department of Statistics, 10 Niepodległości, 61-875 Poznan, Poland, grazyna.dehnel@ue.poznan.pl. ORCiD 0000-0002-0072-9681.

2. An overview of the social cohesion concept

In the EU practice, the level of social cohesion is measured using, among other indicators, the EU Regional Social Progress Index (EU-SPI). The index comprises three dimensions of social progress (Annoni and Dijkstra, 2016, p. 2): basic human needs (nutrition and basic medical care, water and sanitation, shelter (housing), personal safety); foundations of well-being (access to basic knowledge, access to information and communication, health and wellness, environmental quality); opportunity (personal rights, personal freedom and choice, tolerance and inclusion, access to advanced education).

In the article, social cohesion of Polish provinces was analysed on the basis of secondary data. For this reason, the use of the SPI index, which is based on three dimensions (basic human needs, foundations of well-being, opportunity), is considered justified. The final set of variables, used in the study, was selected by the authors of the article (Dehnel et al., 2018). Given this 3-dimensional frame of reference, social cohesion of the Polish provinces was measured using 26 metric variables:

1. **Basic human needs** (7 variables): mean monthly wage (in PLN) – a stimulant, total unemployment rate in % – a destimulant, mean useful floor area of a dwelling per inhabitant in m^2 – a stimulant, average number of persons per room – a destimulant, length of the sewerage network in relation to the length of the water supply network in % – a stimulant, number of doctors and dentists per 10,000 population – a stimulant, crimes reported (criminal offenses, against life and health, against property) per 10,000 population – a destimulant.
2. **Foundations of well-being** (11 variables): people using water treatment services (% of total population) – a stimulant, percentage of all dwellings equipped with central heating – a stimulant, children enrolled in day-care centres per 1000 children up to the age of 3 – a stimulant, children enrolled in nursery schools per 1000 children aged 3–5 – a stimulant, pupils taking obligatory classes of English in primary and intermediate schools (% of all pupils) – a stimulant, number of pupils in secondary schools per class – a destimulant, members of sports club per 1000 population – a stimulant, users of public libraries per 1000 population – stimulant, people participating in cultural events (organised by cultural centres and clubs) per 1000 population – a stimulant, area of public greenspace (parks, residential greenspace) per 10,000 population (in ha) – a stimulant, length of municipal and district improved hard surface roads per 10,000 population (in km) – a stimulant.
3. **Opportunities** (8 variables): persons in households (below the income threshold) using social assistance per 1000 population – a destimulant, age dependency ratio (number of people aged 0–14 and those aged 65 and older per 100 people of working age) – a destimulant, share of women in the labour force in % – a nominant (with the nominal value of 50%), share of youth (up to the age of 25) in the population of registered unemployed in % – a destimulant, share of long-term unemployed (over 12 months) in the population of registered unemployed in % – a destimulant, number of job offers for disabled people per 1000 registered disabled unemployed – a stimulant, places in stationary social welfare facilities per 10,000 population – a stimulant, voter turnout local elections (for municipal authorities and town councils with district rights) in 2014 in % – a stimulant.

The statistical data come from the Local Data Bank maintained by the Central Statistical Office. The reference year is 2016, except for variable "Voter turnout in local elections", which represents data for 2014 (the last local government elections). The x4 nominant variable was converted into a stimulant. The definitions of stimulant, destimulant and nominant can be found in (Walesiak, 2016, p. 18).

3. Social cohesion of Polish provinces – research methodology

The objects analysed in the study were ranked in terms of social cohesion using a two-step procedure, which makes it possible to visualise results of linear ordering. In the first step the objects of interest undergo MDS, as a result of which they can be visualised in a two-dimensional space. In the second step the objects are linearly ordered to produce a ranking. A description on the procedure can be found in (Walesiak, 2016; Walesiak and Dehnel, 2018).

Datasets used in comparative analysis

A ranking of Polish provinces in terms of social cohesion can be obtained on the basis of metric or interval-valued data. For metric data, an observation for the *j*-th variable for the *i*-th object is expressed as a real number. In the case of interval-valued data, observations for each variable are expressed as intervals $x_{ij} = [x_{ij}^l, x_{ij}^u]$ ($x_{ij}^l \leqslant x_{ij}^u$, x_{ij}^l (x_{ij}^u) denotes the lower bound (the upper bound) of the interval). Studies by (Gioia and Lauro, 2006; Brito et al., 2015) provide different examples of data that in real life are of interval type. In this article we compare two approaches to the assessment of social cohesion in Polish provinces:

1. A classical one-step approach, based on metric data, where the ranking of provinces was created using a matrix consisting of 17 objects (16 provinces plus an average province) described by 26 metric variables.

2. A two-stage approach, based on interval-valued data. Firstly, atomic metric data on social cohesion in Polish districts (LAU units) were collected (380 districts described by 26 variables), which were then aggregated to produce interval-valued data. The lower and upper limit of the interval for each province was determined on the basis of district-level data: the minimum and maximum (100% of observations), 1st and 9th deciles (80% of observations) and 1st (Q_1) and 3rd (Q_3) quartiles (50% of observations). Variable values greater than $\left(Q_3 + 3\frac{Q_3-Q_1}{2}\right)$ and less than $\left(Q_3 - 3\frac{Q_3-Q_1}{2}\right)$ are considered outliers. The decision on the selection of the percentage of the cut-off of outlier observations (quartiles and deciles) was made arbitrarily.

The selection of the optimum multidimensional scaling (MDS) procedure

The problem of selecting an optimum MDS procedure is discussed in (Walesiak and Dudek, 2017). Multidimensional scaling was conducted using the smacofSym function from the smacof R package (Mair et al., 2018). To solve the problem of choosing the optimal MDS procedure two criteria were applied in mdsOpt package (Walesiak and Dudek, 2018b): Kruskal's *Stress*-1 fit measure and the Hirschman-Herfindahl *HHI* index, calculated based on Stress per point val-

ues. For all MDS procedures, for which *Stress*-1$_p \leq$ *critical stress*, we choose the one for each occurs min$_p$ $\{HHI_p\}$ (p – MDS procedure number).

For metric data, the optimal MDS procedure was selected after testing 6 normalisation methods (n1, n2, n3, n5, n5a, n12a – see Walesiak and Dudek, 2018a), 5 distance measures (Manhattan, Euclidean, Squared Euclidean, Chebyshev, GDM1[3] – see e.g. Everitt et al., 2011, pp. 49–50), 4 MDS models (ratio, interval, second and third degree polynomial – see Borg and Groenen, 2005; Borg et al., 2018), producing a total of 120 MDS procedures. After applying the optSmacofSym_mMDS function from the mdsOpt package for the R program (R Core Team, 2018) the optimal procedure of MDS was selected, which involves positional standardization (n2), the ratio scaling model and the Manhattan distance.

For interval-valued data, the optimal MDS procedure was selected after testing 6 normalisation methods (n1, n2, n3, n5, n5a, n12a), 4 distance measures (Ichino-Yaguchi, Euclidean Ichino-Yaguchi, Hausdorff, Euclidean Hausdorff – see Billard and Diday, 2006; Ichino and Yaguchi, 1994), 4 MDS models (ratio, interval, second and third degree polynomial), resulting in a total of 96 MDS procedures. After applying the optSmacofSymInterval function from the mdsOpt R package three optimal procedures of MDS were selected for three types of interval-valued data: (a) for an interval between the min and max values: positional standardisation (n2), the ratio scaling model and the Euclidean Ichino-Yaguchi distance, (b) for an interval between 1st and 9th deciles: positional standardisation (n2), mspline 3 scaling model and the Euclidean Ichino-Yaguchi distance, (c) for an interval between 1st and 3rd quartiles: positional standardisation (n2), the ratio scaling model and the Euclidean Hausdorff distance.

After applying the MDS procedures for metric data and three types of interval-valued data, the following results are shown in Figure 1. In each diagram, the anti-pattern (AP) object and the pattern (P) object are connected by a straight line, known as the set axis. Then isoquants of development (curves of equal development) are identified. Objects located between pairs of isoquants represent a similar level of development. Objects located at different points within the same isoquant of development can have the same level of development (as a result of different configurations of variable values).

Ranking of provinces in terms of social cohesion

Based on the results of multidimensional scaling in a two-dimensional space, the provinces can be ranked in terms of social cohesion. Objects are ordered linearly using an aggregated measure (composite indicator) d_i (Hellwig, 1981):

$$d_i = 1 - \sqrt{\sum_{j=1}^{2}(v_{ij} - v_{+j})^2} / \sqrt{\sum_{j=1}^{2}(v_{+j} - v_{-j})^2}, \tag{1}$$

v_{ij} – j-th coordinate for i-th object in a two-dimensional MDS space, v_{+j} (v_{-j}) – j-th coordinate for the pattern (anti-pattern) object in a two-dimensional MDS space.

[3] See Jajuga et al., 2003.

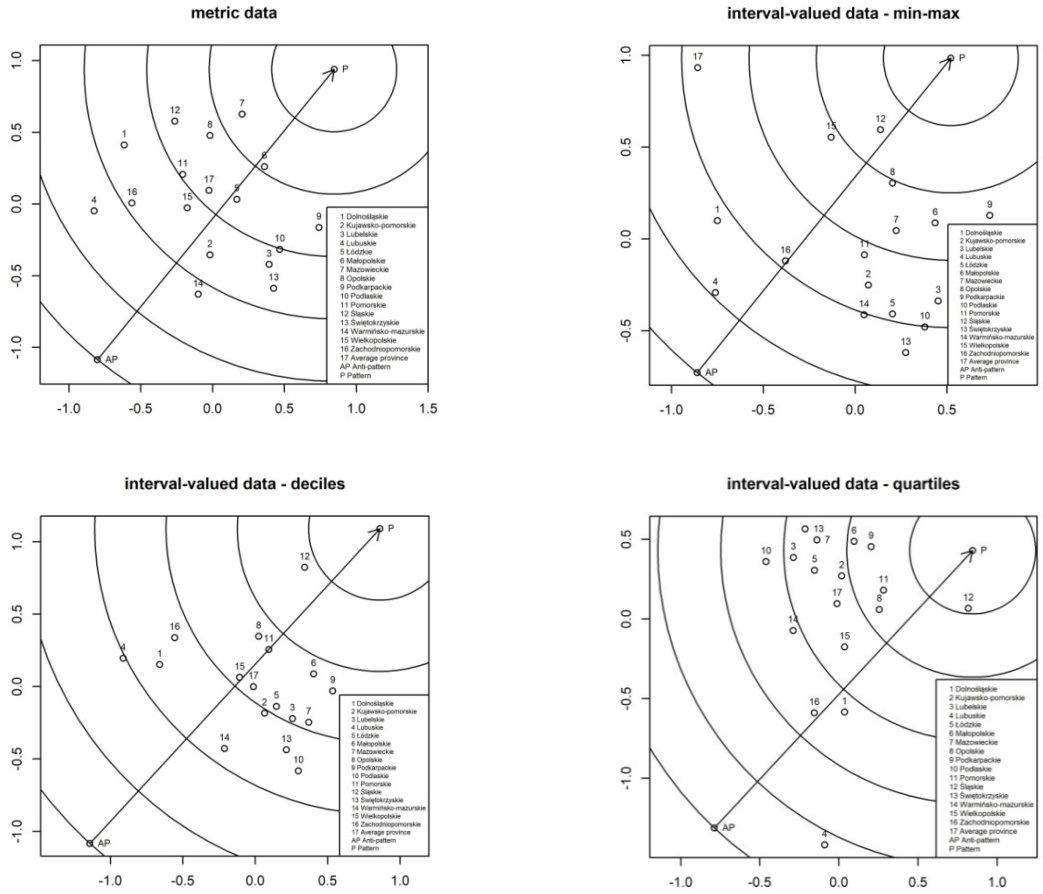

Fig. 1. Results of multidimensional scaling of Polish provinces according to social cohesion

Values of the aggregate measure d_i are included in the interval [0; 1] . The higher the value of d_i, the higher the level of social cohesion of the objects of interest. Table 1 shows the ranking of provinces in terms of social cohesion for 2016.

4. Comparative analysis of the results

The article describes a comparative analysis of rankings (Table 1) of Polish provinces in terms of social cohesion based on metric and interval-valued data between 1st and 3rd quartiles (50% of observations), 1st and 9th deciles (80% of observations) and the minimum and maximum (100% observations).

The comparative analysis was conducted using cluster analysis to identify similarities and differences in the rankings taking the following steps:

1. The four datasets (metric, min-max, deciles, and quartiles) are linearly ordered to produce 4 rankings (see Table 1).
2. The rankings are compared on the basis of Kendall's tau coefficient (Kendall, 1955). For purposes of cluster analysis, values in the rankings are transformed into distances $d = \frac{1}{2}(1 - tau)$ arranged in the form of a distance matrix.

Table 1. Ranking of Polish provinces in terms of social cohesion, based on metric and interval-valued data (3 intervals) for 2016 (values of aggregate measure d_i).

Province	Metric data		min-max		deciles		quartiles	
	d_i	Rank	d_i	Rank	d_i	Rank	d_i	Rank
Dolnośląskie	0.4051	12	0.2961	15	0.3952	15	0.4565	14
Kujawsko-pomorskie	0.4042	13	0.4025	8	0.4916	11	0.6472	6
Lubelskie	0.4507	11	0.3984	9	0.5114	10	0.5252	12
Lubuskie	0.2570	17	0.1777	17	0.3284	17	0.1323	17
Łódzkie	0.5667	5	0.3506	12	0.5198	8	0.5784	9
Małopolskie	0.6806	2	0.5902	5	0.6271	2	0.6856	5
Mazowieckie	0.7269	1	0.5521	6	0.5182	9	0.5869	8
Opolskie	0.6251	3	0.6589	2	0.6214	3	0.7082	4
Podkarpackie	0.5755	4	0.5989	4	0.6052	5	0.7313	3
Podlaskie	0.4975	9	0.3319	13	0.4031	14	0.4524	15
Pomorskie	0.5086	8	0.4685	7	0.6167	4	0.7420	2
Śląskie	0.5534	6	0.7517	1	0.8024	1	0.8473	1
Świętokrzyskie	0.3934	14	0.2629	16	0.4390	13	0.5533	11
Warmińsko-mazurskie	0.2983	16	0.3295	14	0.3713	16	0.4806	13
Wielkopolskie	0.4614	10	0.6452	3	0.5225	7	0.5766	10
Zachodniopomorskie	0.3534	15	0.3522	11	0.4574	12	0.4018	16
Average province	0.5356	7	0.3730	10	0.5275	6	0.6150	7

3. The distance matrix is the basis for cluster analysis, which is conducted using the farthest neighbour method of hierarchical clustering, which can be visualised as a dendrogram (Fig. 2).

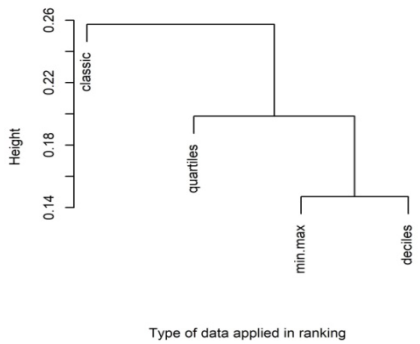

Fig. 2. Dendrogram of rankings of Polish provinces based on 4 types of data

Results of the approach based on interval-valued data are considerably different from those obtained using metric data. The differences increase along with the width of the intervals.

The next analysis focused on the similarity between rankings created with respect to different variables and the ranking based on the aggregate measure, for the four datasets. The following procedure was adopted:

1. The four datasets (metric, min-max, deciles, and quartiles) are linearly ordered according to a set of m variables to produce 4 rankings (see Table 1).
2. For each variable ($j = 1, ..., m$) a distance between each object and the pattern object is calculated according to the formula:
   a) for metric data:

$$d_i = 1 - |x_{ij} - x_{+j}|/(x_{+j} - x_{-j}), \quad j = 1,\ldots, m, \tag{2}$$

   where: $j = 1,\ldots, m$ – variable number, x_{ij} – value of j-th variable for i-th object, x_{+j} (x_{-j}) – j-th coordinate of the pattern (anti-pattern) object.
   b) for interval-valued data (Ichino-Yaguchi distance for one variable):

$$d_i = 1 - |\varphi(x_{ij}, x_{+j})|/\varphi(x_{+j}, x_{-j}) \tag{3}$$

   where: $x_{ij} = [x_{ij}^l, x_{ij}^u]\,(x_{ij}^l \leqslant x_{ij}^u)$ – interval (min-max, 1st and 9th deciles, 1st and 3rd quartiles), $\varphi(x_{ij}, x_{+j}) = |x_{ij} \oplus x_{+j}| - |x_{ij} \otimes x_{+j}| + 0.5(2 \cdot |x_{ij} \otimes x_{+j}| - |x_{ij}| - |x_{+j}|)$, $|\,|$ – interval length, $x_{ij} \oplus x_{+j} = x_{ij} \cup x_{+j}$; $x_{ij} \otimes x_{+j} = x_{ij} \cap x_{+j}$, x_{+j} (x_{-j}) – the pattern (anti-pattern) object for j-th variable.
   It yields a set of m rankings.
3. The general ranking (step 1) is compared with individual rankings (step 2) using Kendall's tau coefficient.
4. Results obtained in step 3 are averaged (see Table 2). A higher average value indicates a higher degree of similarity between the ranking of objects with respect to a given set of variables and the ranking obtained on the basis of the aggregate measure.

Table 2. Assessment of the similarity of rankings of objects with respect to a given set of variables and the ranking obtained on the basis of the aggregate measure

No.	Types of data	Average value of Kendall's tau	Rank
1	Atomic (metric) data	0.0742	4
2	Interval-valued (min-max)	0.1074	2
3	**Interval-valued (1st and 9th deciles)**	**0.1412**	**1**
4	Interval-valued (1st and 3rd quartiles)	0.0861	3

The highest degree of similarity between the rankings of objects with respect to different variables and the ranking obtained on the basis of the aggregate measure is achieved when the intervals are defined by deciles. The decile-based approach can be classified as a robust approach since it reduces the impact of outliers.

Conclusions

The level of social cohesion in Polish provinces was assessed using two approaches: a classical one, based on average metric values, and a symbolic one, based on interval-valued data (min-max, deciles, quartiles). The proposed approach has made it possible to assess social cohesion in provinces not only on the basis of mean values of the variables, but also by taking into account the intervals.

The results of the interval-based approach are considerably different from those obtained using the classical approach (see the dendrogram in Fig. 2). The highest degree of similarity between the rankings of objects with respect to different variables and the ranking obtained on the basis of the aggregate measure is achieved when the intervals are defined by deciles. This approach helps to eliminate the influence of outliers on the assessment of social cohesion in the provinces of Poland.

All the calculations were conducted using scripts written by the authors in the R program.

Acknowledgements

The project is financed by the Polish National Science Centre DEC-2015/17/B/HS4/00905.

References

Annoni, P., & Dijkstra, L. (2016). The EU Regional Social Progress Index: Methodological Note. Brussels: European Commission.

Billard, L., & Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Chichester: John Wiley.

Borg, I., & Groenen, P.J.F. (2005). *Modern Multidimensional Scaling. Theory and Applications*. New York: Springer Science+Business Media.

Borg, I., Groenen, P.J.F., & Mair, P. (2018). *Applied Multidimensional Scaling and Unfolding*. Heidelberg, New York, Dordrecht, London: Springer.

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.

Chan, J., To, H., & Chan, E. (2006). Reconsidering social cohesion: Developing a definition and analytical framework for empirical research. *Social Indicators Research*, 75, 273–302.

Dehnel, G., Walesiak, M., & Obrębalski, M. (2018). Comparative analysis of the ordering of Polish provinces in terms of social cohesion. *Argumenta Oeconomica Cracoviensia* (in press).

Dickes, P., & Valentova, M. (2013). Construction, Validation and Application of the Measurement of Social Cohesion in 47 European Countries and Regions, *Social Indicators Research*, 113, 827–846.

Dickes, P., Valentova, M., & Borsenberger, M. (2010). Construct Validation and Application of a Common Measure of Social Cohesion in 33 European Countries. *Social Indicators Research*, 98, 451–473.

Everitt, B.S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*. Chichester: Wiley.

Hellwig, Z. (1981). *Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych*. In: Welfe, W. (Ed.), *Metody i modele ekonomiczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną*, 46–68. Warszawa: PWE.

Gioia, F., & Lauro, C.N. (2006). Principal component analysis on interval data. *Computational Statistics*, 21(2), 343–363.

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.

Jajuga, K., Walesiak, M., & Bąk, A. (2003). *On the general distance measure,* In: Schwaiger, M., Opitz, O. (Eds.), *Exploratory data analysis in empirical research*, 104–109. Berlin, Heidelberg: Springer-Verlag.

Kendall, M.G. 1955. *Rank correlation methods*. London: Griffin.

Mair, P., De Leeuw, J., Borg, I., & Groenen, P. J. F. (2018). smacof: Multidimensional Scaling, R package ver. 1.10–8. Https://CRAN.R-project.org/package=smacof.

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Https://www.R-project.org

Walesiak, M. (2016). Visualization of linear ordering results for metric data with the application of multidimensional scaling. *Ekonometria*, 2(52), 9–21.

Walesiak, M., & Dehnel, G. (2018). Evaluation of Economic Efficiency of Small Manufacturing Enterprises in Districts of Wielkopolska Province Using Interval-Valued Symbolic Data and the Hybrid Approach. In: Papież, M. and Śmiech, S. (Eds.), *The 12th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena*. Conference Proceedings, Foundation of the Cracow University of Economics, Cracow, 563–572.

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. (2018a). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. *R package*, version 0.47-3. http://CRAN.R-project.org/package=clusterSim

Walesiak, M., & Dudek, A. (2018b). mdsOpt: Searching for Optimal MDS Procedure for Metric Data. *R package*, version 0.3-3. http://CRAN.R-project.org/package=mdsOpt