STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 693–710, DOI 10.21307/stattrans-2018-036

THE CHOICE OF NORMALIZATION METHOD AND RANKINGS OF THE SET OF OBJECTS BASED ON COMPOSITE INDICATOR VALUES

Marek Walesiak[1]

ABSTRACT

The choice of the normalization method is one of the steps for constructing a composite indicator for metric data (see, e.g. Nardo et al., 2008, pp. 19-21). Normalization methods lead to different rankings of the set of objects based on composite indicator values. In the article 18 normalization methods and 5 aggregation measures (composite indicators) were taken into account. In the first step the groups of normalization methods, leading to identical rankings of the set of objects, were identified. The considerations included in Table 3 reduce this number to 10 normalization methods. Next, the article discusses the procedure which allows separating groups of normalization methods leading to similar rankings of the set of objects separately for each composite indicator formula. The proposal, based on Kendall's tau correlation coefficient (Kendall, 1955) and cluster analysis, can reduce the problem of choosing the normalization method. Based on the suggested research procedure the simulation results for five composite indicators and ten normalization methods were presented. Moreover, the proposed approach was illustrated by an empirical example. Based on the analysis of the dendrograms three groups of normalization methods were separated. The biggest differences in the results of linear ordering refer to methods n2, n9a against the other normalization methods.

Key words: variables normalization, rankings, composite indicators, Kendall's tau correlation coefficient, cluster analysis.

1. Introduction

Simulation studies, allowing the alignment of linear ordering (ranking) of the set of objects, via composite indicators values, procedures (the procedure takes into account weights of variables, selected normalization methods and selected constructions of aggregation measures), from the perspective of determining the correctness (quality) of aggregated variables, were conducted by (Grabiński, 1984) and (Bąk, 1999). T. Grabiński (see Grabiński, 1984, pp. 58–62; Grabiński,

[1] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra. E-mail: marek.walesiak@ue.wroc.pl.

Wydymus and Zeliaś, 1989, pp. 122–123) suggested five groups of correctness (quality) measures for determining the value of aggregated variables:

1. Measures of compatibility of distance matrices calculated for objects in m-dimensional space of the variables and 1-dimensional space of the aggregated variable (3 measures).
2. Measures based on Pearson product-moment correlation coefficient between m variables and the aggregated variable (2 measures).
3. Measures based on Spearman's rank correlation coefficient between *m* variables and the aggregated variable (3 measures).
4. Measures determining an average taxonomic distance of the aggregated variable from the m variables (2 measures).
5. Measures characterizing the variability level and concentration for the aggregated variable (2 measures).

The better a given linear ordering procedure (taking into account weights of variables, selected normalization methods and selected constructions of aggregation measures), the lower are the values of these measures (Grabiński, Wydymus and Zeliaś, 1989, p. 125). The author does not justify substantively the introduced measures. The doubts related to their application are presented based on two groups of measures.

The first group of measures covers the selected compatibility functions applied in multidimensional scaling (e.g. STRESS-1 function – see Borg and Groenen, 2005, p. 42). Based on the distance matrix between objects in m-dimensional space, such mapping of the set of objects into a set of points in r-dimensional space is sought ($r < m$, in linear ordering $r = 1$ – of the aggregated variable values), which allows achieving the best possible compatibility. The objects distant from each other in m-dimensional space shall also remain distant in *r*-dimensional space (1-dimensional). The situation is, however, different in the case of linear ordering. A distant object, from the perspective of the initial set of *m* variables, can be found at the same distance from the pattern object. Therefore, the distance between them, in terms of the aggregated variable, may equal zero.

In the second group, e.g. the measure of linear correlation of the aggregated variable with diagnostic variables was suggested, which takes the following form (the so-called uncertainty coefficient):

$$M_4 = 1 - \frac{1}{m}\sum_{j=1}^{m} r_{\cdot j}, \tag{1}$$

where:$r_{\cdot j}$ – linear correlation coefficient for *j*-th variable with the aggregated variable,

$j = 1, \dots, m$ – variable number.

The most preferred value of this measure is 0, when all correlation coefficients of diagnostic variables with the aggregated variable equal 1. Such approach is missing substantive justification in the case of linear ordering.

Due to an ambiguous interpretation of correctness (quality) measures for determining values of aggregated variables a different approach was used in the article.

There is a growing demand for various rankings of the set of objects (e.g. countries, regions) due to, for example, their competitiveness, tourist attractiveness, social cohesion, socio-economic development, and environmental pollution. Analyses using aggregate measures require normalization of variable values. Normalization methods lead to different rankings of the set of objects based on aggregation measures (composite indicators) values. The proposed approach allows to objectivize the results of analyses in this area.

In the article 18 normalization methods and 5 aggregation measures (composite indicators) were taken into account. Two elements of the article should be considered innovative:

- identification of groups of normalization methods resulting in identical values and identical orderings for the aggregation measures (see Table 3),
- the proposal of the procedure allowing the separation of the groups of normalization methods leading to similar rankings of the set of objects (see section 3).

The proposal, based on Kendall's tau correlation coefficient and cluster analysis, can reduce the problem of choosing the normalization method. Based on the suggested research procedure the simulation results were presented. Moreover, the proposed approach was illustrated by an empirical example.

2. Steps for Constructing a Composite Indicator

The general procedure in linear ordering (ranking) of the set of objects via composite indicators values, carried out based on metric data (measured on an interval scale and ratio scale)[2], takes the following form (see Grabiński, Wydymus and Zeliaś, 1989, p. 92; Pawełek, 2008, pp. 110–111; Nardo et al., 2008, pp. 19–21):

a) for methods based on pattern object (there are two types of pattern objects: upper pattern – ideal object, upper pole, lower pattern – anti-ideal object, lower pole):

$$P \rightarrow A \rightarrow X \rightarrow [x_{ij}] \rightarrow SDN \rightarrow T_w \rightarrow N \rightarrow SM_w \rightarrow R, \qquad (2)$$

where:

P – choice of a complex phenomenon (the overriding multidimensional phenomenon for ordering A set elements, which is not subject to direct measurement),

A – choice of objects,

X – selection of variables,

$[x_{ij}]$ – collecting data and the construction of data matrix (x_{ij} – value for *j*-th variable for *i*-th object),

SDN – identifying preferential variables (stimulant, destimulant, nominant). M_j variable is a stimulant (see Hellwig, 1981, p. 48), when for every two of its observations x_{ij}^S, x_{kj}^S referring to objects A_i, A_k, it takes $x_{ij}^S > x_{kj}^S \Rightarrow A_i \succ A_k$ ($\succ$ means A_i object domination over A_k object). M_j variable is a destimulant (see Hellwig, 1981, p. 48), when for every two of its observations x_{ij}^D, x_{kj}^D referring to objects A_i, A_k take $x_{ij}^D > x_{kj}^D \Rightarrow A_i \prec A_k$ ($\prec$ means A_k object domination over

[2] The characteristics of measurement scales is presented, e.g. in the studies by (Stevens, 1946; Walesiak, 1995; Walesiak, 2011, pp. 13–16).

A_i object). Therefore, M_j variable represents a unimodal nominant (see Borys, 1984, p. 118), when for every two of its observations x_{ij}^N, x_{kj}^N referring to objects A_i, A_k (nom_j means the nominal level of j-th variable) if $x_{ij}^N, x_{kj}^N \leq nom_j$, then $x_{ij}^N > x_{kj}^N \Rightarrow A_i \succ A_k$; if $x_{ij}^N, x_{kj}^N > nom_j$, then $x_{ij}^N > x_{kj}^N \Rightarrow A_i \prec A_k$,

T_w – transformation of nominants into stimulants (required for an anti-ideal object only). Transformation formulas can be found, e.g. in the study by (Walesiak, 2011, p. 18),

N – normalization of variable values,

SM_w – composite indicator calculation by aggregating normalized variables – the application of distance measures from pattern object using weights. The coordinates of upper pattern object covers the most preferred variable values (maximum for a stimulant, minimum for a destimulant). The coordinates of lower pattern object cover the least preferred variable values (minimum for a stimulant, maximum for a destimulant),

R – ordering of objects (ranking) in accordance with the composite indicator values.

b) for methods not based on pattern object:

$$P \rightarrow A \rightarrow X \rightarrow [x_{ij}] \rightarrow SDN \rightarrow T_b \rightarrow N \rightarrow SM_b \rightarrow R, \tag{3}$$

where:

T_b – transformation of destimulants and nominants into stimulants. Transformation formulas are presented, e.g. in the study by (Walesiak, 2011, p. 18),

SM_b – composite indicator calculation by aggregating normalized variables – averaging normalized variable values using weights.

In linear ordering, carried out based on metrical data, the choice of the normalization method for variable values remains one of the stages. The purpose of normalization is to adjust the size (magnitude) and the relative weighting of the input variables (see, e.g. Milligan and Cooper, 1988, p. 182). An overview of normalization methods for variable values is presented in the study by (Walesiak, 2014b). Table 1 presents normalization methods of linear transformation (see e.g. Jajuga and Walesiak, 2000, pp. 106–107; Zeliaś, 2002, p. 792):

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

Table 1. Normalization methods

Type	Method	Parameter		Measurement scale of variables	
		B_j	A_j	before normalization	after normalization
n1	Standardization	s_j	$\bar{x}_j$	ratio or interval	Interval
n2	Positional standardization	mad_j	med_j	ratio or interval	Interval
n3	Unitization	r_j	$\bar{x}_j$	ratio or interval	Interval
n3a	Positional unitization	r_j	med_j	ratio or interval	Interval

Table 1. Normalization methods (cont.)

n4	Unitization with zero minimum	r_j	$\min\limits_i\{x_{ij}\}$	ratio or interval	Interval
n5	Normalization in range [–1; 1]	$\max\limits_i\lvert x_{ij}-\bar{x}_j\rvert$	$\bar{x}_j$	ratio or interval	Interval
n5a	Positional normalization in range [–1; 1]	$\max\limits_i\lvert x_{ij}-med_j\rvert$	med_j	ratio or interval	Interval
n6	Quotient transformations	s_j	0	ratio	Ratio
n6a		mad_j	0	ratio	Ratio
n7		r_j	0	ratio	Ratio
n8		$\max\limits_i\{x_{ij}\}$	0	ratio	Ratio
n9		$\bar{x}_j$	0	ratio	Ratio
n9a		med_j	0	ratio	Ratio
n10		$\sum_{i=1}^{n} x_{ij}$	0	ratio	Ratio
n11		$\sqrt{\sum_{i=1}^{n} x_{ij}^2}$	0	ratio	Ratio
n12	Normalization	$\sqrt{\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}$	$\bar{x}_j$	ratio or interval	Interval
n12a	Positional normalization	$\sqrt{\sum_{i=1}^{n}\left(x_{ij}-med_j\right)^2}$	med_j	ratio or interval	Interval
n13	Normalization with zero being the central point	$\frac{r_j}{2}$	m_j	ratio or interval	Interval

$\bar{x}_j$ – mean for *j*-th variable, s_j – standard deviation for *j*-th variable, r_j – range for *j*-th variable, $m_j = \frac{\max\limits_i\{x_{ij}\}+\min\limits_i\{x_{ij}\}}{2}$ – mid-range for *j*-th variable, $med_j = \operatorname*{med}\limits_i\left(x_{ij}\right)$ – median for *j*-th variable, $mad_j = \operatorname*{mad}\limits_i\left(x_{ij}\right)$ – median absolute deviation for *j*-th variable.

Source: (Walesiak, 2014b, pp. 364–365).

where:

x_{ij} – value for *j*-th variable for *i*-th object,

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

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

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

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

Column 1 in Table 1 presents the type of normalization formula adopted as the function data.Normalization of clusterSim package (see Walesiak and Dudek, 2018) of R program (R Core Team, 2018).

An aggregation measure SM_i represents the tool for linear ordering methods as a sub-function aggregating partial information contained in particular variables and determined for each object from the set of objects. Generally, the constructions of aggregation measures (composite indicators) can be divided as follows (cf. e.g. Grabiński 1984, p. 38):

– based on pattern object (e.g. Hellwig's measure of development; GDM1 distance; TOPSIS measure),

– not based on pattern object (arithmetic mean, harmonic mean, geometric mean; median).

Table 2 presents five constructions of aggregation measures (composite indicators) (four based on pattern object ones to be followed by one not based on pattern object) applied for metric data to be used later in the article.

Table 2. Constructions of aggregation measures (composite indicators) used for linear ordering (ranking) of objects

No.	Name	SM_i
1	GDM1 distance (Walesiak, 2002; Jajuga, Walesiak and Bąk, 2003)	$1 - GDM1_i^+ = \frac{1}{2} + \frac{\sum_{j=1}^{m} \alpha_j (z_{ij} - z_{wj})(z_{wj} - z_{ij}) + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,w}}^{n} \alpha_j (z_{ij} - z_{lj})(z_{wj} - z_{lj})}{2 \left[\sum_{j=1}^{m} \sum_{l=1}^{n} \alpha_j (z_{ij} - z_{lj})^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} \alpha_j (z_{wj} - z_{lj})^2 \right]^{0,5}}$
2	Measure of development (Hellwig, 1968; 1972)	$1 - \frac{d_{iw}^+}{\bar{d}_{.w}^+ + 2s_d}$
3	TOPSIS measure (Hwang and Yoon, 1981)	$\frac{d_{iw}^-}{d_{iw}^- + d_{iw}^+}$
4	GDM1_TOPSIS – TOPSIS measure with GDM1 distance (Walesiak, 2014a)	$\frac{GDM1_i^-}{GDM1_i^- + GDM1_i^+}$
5	Arithmetic mean	$\sum_{j=1}^{m} \alpha_j z_{ij}$

SM_i – aggregation measure (composite indicator) value for *i*-th object (the resulting aggregate variable has stimulant interpretation), $i, l = 1, \ldots, n$ – object number, w – pattern object number, $j = 1, \ldots, m$ – variable number, z_{wj} – *j*-th coordinate of a pattern object, α_j – weight for *j*-th variable ($\alpha_j \in [0; 1]$ and $\sum_{j=1}^{m} \alpha_j = 1$), $d_{iw} = \sqrt{\sum_{j=1}^{m} \alpha_j^2 (z_{ij} - z_{wj})^2}$ – weighted Euclidean distance for *i*-th object from a pattern object, $GDM1_i^-$ and $GDM1_i^+$ – GDM1 distance for *i*-th object from the lower pole (anti-ideal object) and the upper pole (ideal object), d_{iw}^- and d_{iw}^+ – weighted Euclidean distance for *i*-th object from the lower and upper pole, $\bar{d}_{.w} = \frac{1}{n} \sum_{i=1}^{n} d_{iw}^+$, $s_d = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (d_{iw}^+ - \bar{d}_{.w})^2}$.

Source: Author's compilation.

3. Research Procedure Allowing Separation of the Groups of Normalization Methods Resulting in a Similar Linear Ordering of a Set of Objects

The research procedure allowing separation of the groups of normalization methods for variable values resulting in a similar linear ordering (ranking) of a set of objects covers the following steps:

1. Linear ordering of the set of objects is performed in accordance with the general procedure used in linear ordering methods illustrated in section 2 (scheme (2) or (3)). Any acceptable methods presented in Table 1 are used in the normalization of variable values (for ratio variables 18 normalization methods are possible and for interval variables – 10 normalization methods).
2. Object ordering obtained for the acceptable normalization methods is compared with the application of Kendall's tau correlation coefficient Γ_{rs} (see Kendall and Buckland, 1986, p. 266; Kendall, 1955, p. 19; Walesiak, 2011, pp. 36-38). Kendall's tau correlation coefficient takes values in interval $[-1;1]$. The value of 1 means complete compatibility of orderings, whereas the value -1 implies their complete opposition. For the purposes of cluster analysis, Kendall's tau correlation coefficients are transformed into distances using the following formula:

$$d_{rs} = \frac{1}{2}(1 - \Gamma_{rs}), \tag{5}$$

where:

$d_{rs} \in [0;1]$, $d_{rs} = 0$, when $\Gamma_{rs} = 1$ and $d_{rs} = 1$, when $\Gamma_{rs} = -1$,
r, s – numbers of normalization methods.

3. Cluster analysis is carried out based on the distance matrix $[d_{rs}]$, which allows separating groups of normalization methods for variable values resulting in similar linear ordering of a set of objects. In this case it is possible to use one of many classification methods (see, e.g. Everitt et al., 2011). The agglomerative hierarchical method of the farthest neighbour clustering was applied in the article.

Certain observations can be put forward regarding normalization methods presented in Table 3 for aggregation measures (SM_i) obtained using the following distance measures: GDM1, Hellwig's measure of development, TOPSIS, GDM1_TOPSIS and SM_i taking the form of an arithmetic mean.

Table 3. Groups of normalization methods resulting in identical values and identical orderings for the aggregation measures (SM_i) from Table 2

Groups of methods	Identical SM_i values		Identical orderings (rankings)
	Distances: GDM1 and GDM1_TOPSIS	Hellwig's measure of development, TOPSIS	all SM_i constructions from Table 2
A	n3, n3a, n4, n7, n13	n3, n3a, n4, n7	n3, n3a, n4, n7, n13
B	n1, n6, n12	n1, n6	n1, n6, n12
C	n2, n6a	n2, n6a	n2, n6a
D	n9, n10	–	n9, n10

Source: Author's compilation.

Identical SM_i values (and thus identical orderings) for A, B, C and D groups of formulas in the case of GDM1 and GDM1_TOPSIS distance measures result from the fact that GDM1 distance does not depend on the shift parameter used in normalization methods (4). Furthermore, multiplying normalized values by a constant does not change GDM1 distance:

– for n13 formula the constant equals 2:

$$z_{ij} = \frac{x_{ij}}{r_j/2} - \frac{m_j}{r_j/2} = 2\left(\frac{x_{ij}}{r_j} - \frac{m_j}{r_j}\right), \tag{6}$$

– for n12 formula the constant equals $\sqrt{\frac{1}{n-1}}$:

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}} - \frac{\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}} = \sqrt{\frac{1}{n-1}}\left(\frac{x_{ij}}{s_j} - \frac{\bar{x}_j}{s_j}\right), \tag{7}$$

– for n10 formula the constant equals $1/n$:

$$z_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} = \frac{1}{n}\frac{x_{ij}}{\bar{x}_j}. \tag{8}$$

Identical SM_i values (and thus identical orderings), in the case of Hellwig's measure of development and TOPSIS, result from the fact that Euclidean distance applied in these measures does not depend on the shift parameter used in normalization methods, but only on the scale parameter which is identical for A, B and C groups of methods (see Pawełek, 2008, p. 94).

Additionally, n13 is present in A group of normalization methods, while in group B – n12 formula. Two normalization methods n9 and n10 result in identical object ordering. For n13, n12 and n10 formulas normalized values are multiplied by a constant. This causes a change in Euclidean distance, however, does not change the ordering of objects.

In the case of a aggregation measure, taking the form of an arithmetic mean, identical orderings result from the fact that the shift parameter, used in normalization methods, does not change the order of objects (in fact a constant is subtracted from SM_i value of each object). Multiplying SM_i value by a constant does not alter the order of objects either. For example, for n1, n6 and n12 formulas from group B the following is obtained:

$$\text{for n1: } SM_i = \sum_{j=1}^{m}\left(\frac{x_{ij}}{s_j} - \frac{\bar{x}_j}{s_j}\right) = \sum_{j=1}^{m}\frac{x_{ij}}{s_j} - \sum_{j=1}^{m}\frac{\bar{x}_j}{s_j}, \tag{9}$$

$$\text{for n6: } SM_i = \sum_{j=1}^{m}\frac{x_{ij}}{s_j}, \tag{10}$$

$$\text{for n12: } SM_i = \sum_{j=1}^{m}\left(\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}} - \frac{\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}\right) = \sqrt{\frac{1}{n-1}}\left(\sum_{j=1}^{m}\frac{x_{ij}}{s_j} - \sum_{j=1}^{m}\frac{\bar{x}_j}{s_j}\right). \tag{11}$$

The order of objects, determined in line with n6 normalization method, does not change n1 formula (subtracting a constant from each SM_i value obtained for n6 formula) and for n12 formula (here subtracting a constant takes place and next multiplying by a different constant).

4. The Results of Simulation Analyses

The research procedure discussed in section 3 was used in simulation analyses, which allows separating groups of normalization methods for variable values resulting in similar linear orderings of a set of objects using a specific aggregation measure (SM_i):

1. Multivariate normal distribution was used to generate data (function rmnorm from the mnormt package – see Genz and Azzalini, 2016) based on models presented in Table 4. A simplifying assumption was adopted that the set of analysed variables includes stimulants only. Correlation with aggregate variable (vector of SM_i values) is positive for stimulants (see Grabiński, 1992, p. 138). Due to the transitivity of variables correlation[3] (Hellwig, 1976) it was adopted that the correlation between stimulants will also be positive. Therefore, models in Table 4 take values of correlation coefficients from 0.2 to 0.95 between variables in data matrix. The generated data differ in terms of variables' order of magnitude (see mean values for variables) and the variability measured by the coefficient of variation (0.20, 0.16, 0.24, 0.10).

Table 4. The characteristics of models in simulation analysis

No.	Mean values for variables	Covariance matrix Σ	Correlation matrix $[r_{jl}]$
1	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 14 & 42 & 70 \\ 14 & 400 & 420 & 700 \\ 42 & 420 & 3600 & 2100 \\ 70 & 700 & 2100 & 10000 \end{bmatrix}$	$r_{jj} = 1, r_{jl} = 0.35$ $1 \le j, l \le 4$
2	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 26 & 78 & 130 \\ 26 & 400 & 780 & 1300 \\ 78 & 780 & 3600 & 3900 \\ 130 & 1300 & 3900 & 10000 \end{bmatrix}$	$r_{jj} = 1, r_{jl} = 0.65$ $1 \le j, l \le 4$
3	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 38 & 114 & 190 \\ 38 & 400 & 1140 & 1900 \\ 114 & 1140 & 3600 & 5700 \\ 190 & 1900 & 5700 & 10000 \end{bmatrix}$	$r_{jj} = 1, r_{jl} = 0.95$ $1 \le j, l \le 4$
4	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 36 & 90 & 120 \\ 36 & 400 & 1080 & 1000 \\ 90 & 1080 & 3600 & 3600 \\ 120 & 1000 & 3600 & 10000 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.9 & 0.75 & 0.6 \\ 0.9 & 1 & 0.9 & 0.5 \\ 0.75 & 0.9 & 1 & 0.6 \\ 0.6 & 0.5 & 0.6 & 1 \end{bmatrix}$
5	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 8 & 60 & 140 \\ 8 & 400 & 480 & 1200 \\ 60 & 480 & 3600 & 1800 \\ 140 & 1200 & 1800 & 10000 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.2 & 0.5 & 0.7 \\ 0.2 & 1 & 0.4 & 0.6 \\ 0.5 & 0.4 & 1 & 0.3 \\ 0.7 & 0.6 & 0.3 & 1 \end{bmatrix}$

Source: Author's compilation.

2. Normalization of variables was carried out using the methods from Table 1. Due to the fact that the groups of A, B, C and D normalization methods result

[3] Let Y represent the aggregated variable, whereas X_1 and X_2 two variables from the data matrix. For $r_{X_1Y} = 0.9$ and $r_{X_2Y} = 0.95$ correlation coefficient $r_{X_1X_2}$ can only take values in the interval $0.719 \le r_{X_1X_2} \le 0.991$. On the other hand, for $r_{X_1Y} = 0.6$ and $r_{X_2Y} = 0.8$, correlation coefficient $r_{X_1X_2}$ can only take values in the interval $0 \le r_{X_1X_2} \le 0.96$.

in identical ordering, further analysis covered first methods from the indicated groups (n1, n2, n3, n9) and the other methods (n5, n5a, n8, n9a, n11, n12a).

3. Linear ordering was conducted using five aggregation measures (SM_i) listed in Table 2 (equal weights were used for variables).
4. For each individual aggregation measure (SM_i) the ordering of objects was compared by applying 10 normalization methods. Kendall's tau correlation coefficient Γ_{rs} was used to compare the ordering of objects, which gave 10 x 10 matrix.
5. Cluster analysis of normalization methods for variable values was carried out for 10 x 10 matrix. For the purposes of the analysis Kendall's tau correlation coefficient was transformed into distances using formula (5). The agglomerative hierarchical method of the farthest neighbour clustering was applied to separate groups of normalization methods for variable values resulting in similar linear orderings of the set of objects using a specific SM_i.

20 sets of data were generated for each model from Table 4, the procedure was conducted in accordance with points 2-5 divided into 2, 3 and 4 classes and next the obtained classification results of five aggregation measures (SM_i) from Table 2 were compared using the adjusted Rand index (see Hubert and Arabie, 1985). Table 5 presents the outcome of compatibility comparison for cluster analysis results of normalization methods for five aggregation measures (SM_i) taking the mean value of the adjusted Rand index.

Table 5. Compatibility comparison of cluster analysis results of normalization methods for five aggregation measures (SM_i) taking the mean value of the adjusted Rand index

Model 1						Model 2					
	1	2	3	4	5		1	2	3	4	5
1	1.000	0.914	0.886	0.891	0.870	1	1.000	**0.922**	0.841	0.826	0.813
2	**0.914**	1.000	**0.916**	**0.922**	0.865	2	**0.922**	1.000	0.833	0.839	0.834
3	0.886	0.916	1.000	0.890	**0.908**	3	0.841	0.833	1.000	**0.899**	0.824
4	0.891	**0.922**	0.890	1.000	0.859	4	0.826	0.839	**0.899**	1.000	**0.867**
5	0.870	0.865	0.908	0.859	1.000	5	0.813	0.834	0.824	0.867	1.000
Model 3						Model 4					
	1	2	3	4	5		1	2	3	4	5
1	1.000	**0.865**	0.800	0.824	**0.823**	1	1.000	**0.884**	0.885	0.893	0.862
2	**0.865**	1.000	0.801	0.774	0.808	2	0.884	1.000	0.861	0.817	0.844
3	0.800	0.801	1.000	**0.853**	0.806	3	0.885	0.861	1.000	0.892	0.873
4	0.824	0.774	**0.853**	1.000	0.806	4	**0.893**	0.817	**0.892**	1.000	**0.896**
5	0.823	0.808	0.806	0.806	1.000	5	0.862	0.844	0.873	**0.896**	1.000
Model 5						Mean (models 1-5)					
	1	2	3	4	5		1	2	3	4	5
1	1.000	**0.955**	0.879	0.898	0.869	1	1.000	**0.908**	0.858	0.866	0.847
2	**0.955**	1.000	0.870	**0.930**	0.877	2	**0.908**	1.000	0.856	0.856	0.845
3	0.879	0.870	1.000	0.908	0.857	3	0.858	0.856	1.000	**0.888**	0.853
4	0.898	0.930	**0.908**	1.000	**0.878**	4	0.866	0.856	**0.888**	1.000	**0.861**
5	0.869	0.877	0.857	0.878	1.000	5	0.847	0.845	0.853	0.861	1.000

1 – GDM1 distance, 2 – Hellwig's measure of development, 3 – TOPSIS measure, 4 – TOPSIS measure with GDM1 distance, 5 – arithmetic mean. Minimum values are underlined, maximum values are in **bold** (excluding the main diagonal).

Source: Author's compilation using R program.

Having analysed the obtained results of compatibility comparison for cluster analysis of normalization methods for five aggregation measures (SM_i), taking the mean value of the adjusted Rand index, the following conclusions can be drawn:

1. Values of the adjusted Rand index for models 1-5 vary in the interval $[0.774, 0.955]$. Mean values of the adjusted Rand index taken from five models are in the interval $[0.845, 0.908]$. Therefore, the results of cluster analysis of normalization methods for the analysed aggregation measures (SM_i) are similar to each other.
2. Dendrite of cluster analysis results' similarity of normalization methods for five aggregation measures is presented in Figure 1 (developed based on the matrix for models 1-5 from Table 5).

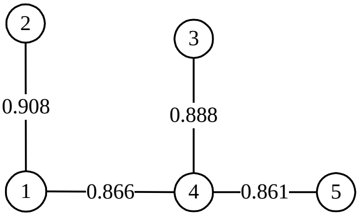


Figure 1. Dendrite of cluster analysis results' similarity of normalization methods for five aggregation measures

1 – GDM1 distance, 2 – Hellwig's measure of development, 3 – TOPSIS measure, 4 – TOPSIS measure with GDM1 distance, 5 – arithmetic mean.

Source: Author's compilation.

5. Empirical Research Results

The evaluation of tourism competitiveness level of the Sudety communes in Poland, covering 52 out of 169 communes in Lower Silesia region, was carried out in the article (Gryszel and Walesiak, 2018). The Sudety communes are located in the geographical area of the Sudety in the southern part of Lower Silesia region. They are characterized by the most valuable tourism advantages, where the tourism function either dominates or is of great importance among other economic functions in a commune. The following variables were used in the study:

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

x2 – beds in other accommodation facilities per 1 km^2 of a commune area,

x3 – number of nights of resident tourists (Poles) falling per day per 1000 inhabitants of a commune,

x4 – number of nights of foreign tourists falling per day per 1000 inhabitants of a commune,

x5 – communal expenditure for tourism per 1000 inhabitants in PLN,

x6 – funds obtained from the European Union and from the state budget to finance the EU programs and projects per 1 inhabitant in PLN, x7 – number of tourist economy entities per 1000 inhabitants of a commune (natural persons conducting economic activity),

x8 – number of tourist economy entities per 1000 inhabitants of a commune (legal persons and organizational entities without legal personality).

All variables represent stimulants. Statistical data were collected in 2012 and retrieved from the Local Data Bank (LDB). The research procedure discussed in section 3 was used in the article, which allows separating the groups of methods for the normalization of variable values, resulting in similar linear ordering of the set of communes in terms of their tourism competitiveness level. The analysed variables are measured using ratio scale, therefore all normalization methods listed in Table 1 are acceptable.

The results of linear ordering compatibility for 52 Sudety communes, in terms of their tourism competitiveness level, using 18 normalization methods and 5 SM_i from Table 2 are presented in Figure 1. Due to the fact that the groups of A, B, C and D normalization methods result in identical orderings, further analysis covered first formulas from the indicated groups (n1, n2, n3, n9) and other formulas (n5, n5a, n8, n9a, n11, n12a).

Regardless of the adopted SM_i construction, the results of linear ordering compatibility for 52 Sudety communes, in terms of their tourism competitiveness level, using 10 normalization methods are analogical in this case. Table 6 contains compatibility comparison of cluster analysis results of normalization methods (after splitting the dendrograms from Figure 2 into 2, 3 and 4 clusters) for five aggregation measures (SM_i) taking the mean value of the adjusted Rand index.

Table 6. Compatibility comparison of cluster analysis results of normalization methods (after splitting the dendrograms into 2, 3 and 4 clusters) for five aggregation measures (SM_i) taking the mean value of the adjusted Rand index.

Aggregation measure	1	2	3	4	5
1	1.0000	1.0000	0.7674	0.7674	0.7674
2	1.0000	1.0000	0.7674	0.7674	0.7674
3	0.7674	0.7674	1.0000	1.0000	1.0000
4	0.7674	0.7674	1.0000	1.0000	1.0000
5	0.7674	0.7674	1.0000	1.0000	1.0000

1 – GDM1 distance, 2 – Hellwig’s measure of development, 3 – TOPSIS measure, 4 – TOPSIS measure with GDM1 distance, 5 – arithmetic mean.

Source: Author’s compilation.

Taking into account results from Table 6 the following conclusions can be drawn:

1. The results of cluster analysis of normalization methods (after splitting the dendrograms into 2, 3 and 4 clusters) for TOPSIS measure, TOPSIS measure with GDM1 distance, and arithmetic mean are the same.
2. The results of cluster analysis of normalization methods (after splitting the dendrograms into 2, 3 and 4 clusters) for GDM1 distance and Hellwig's measure of development are the same.
3. The differences between groups of aggregation methods listed in points 1 and 2 relate to the division of dendrograms into two clusters.

Based on the analysis of the dendrograms in Figure 2 three groups of normalization methods were separated:

group 1 (6 methods): n1, n3, n5, n5a, n8, n12a,

group 2 (2 methods): n2, n9a,

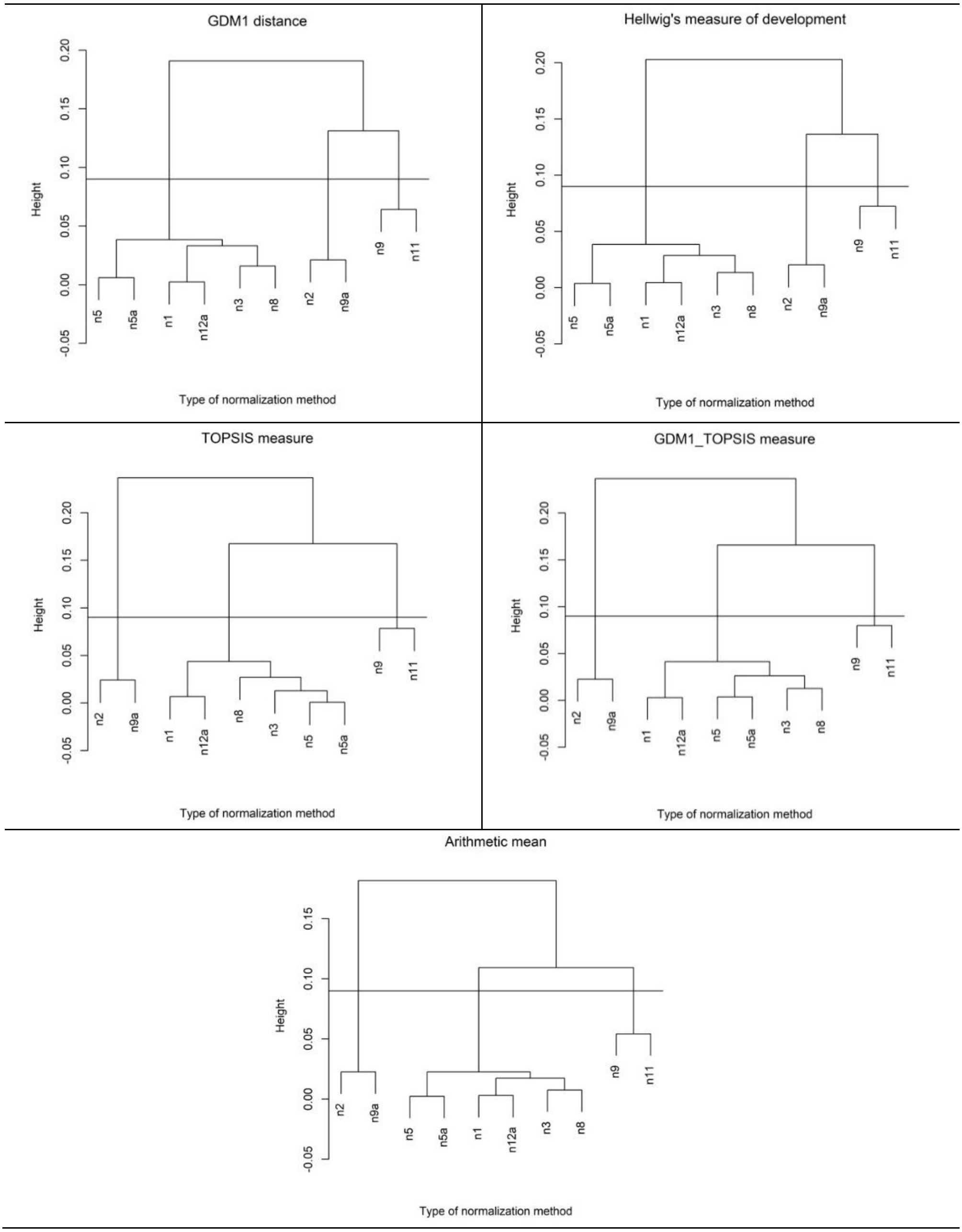
group 3 (2 methods): n9, n11.

The results presented in Figure 2 regarding the adopted SM_i construction differ, for the separated groups of normalization methods, in the level of class links in a dendrogram.

The biggest differences in the results of linear ordering refer to methods n2, n9a against the other normalization methods.

The presented proposal allows reducing the problem of a normalization method selection. Significant differences between the results of linear ordering appear in the analysed case for the normalization methods from different groups.

In the current practice, not considering the proposed procedure, selecting 18 methods for normalizing variable values for metric data, we had 18 proposals to choose from (see Table 1). The considerations included in Table 3 reduce this number to 10 normalization methods. The choice still becomes arbitrary and difficult to justify. The proposed approach does not completely solve the problem, but it allows distinguishing groups of normalization methods leading to similar results of linear ordering (rankings) of objects. In the analysed example, we already have three types of normalization methods to choose from (normalization methods in the same groups give similar results of linear ordering of objects). Therefore, the presented proposal allows limiting the problem of selecting the normalization method of variable values.


GDM1 distance

Hellwig's measure of development

TOPSIS measure

GDM1_TOPSIS measure

Arithmetic mean

Figure 2. The results of linear ordering compatibility for 52 Sudety communes in terms of their tourism competitiveness level using 10 normalization methods and 5 aggregation measures (dendrograms of normalization methods similarity)

Source: Author’s compilation using R program.

6. Conclusions

Normalization methods lead to different rankings of the set of objects based on aggregation measure (composite indicator) values. The study includes 18 normalization methods and 5 aggregation measures (composite indicators).

The groups of normalization methods were indicated, which results in identical SM_i values and identical orderings for SM_i obtained using the following distance measures: GDM1, Hellwig's measure of development, TOPSIS, GDM1_TOPSIS and aggregation measure (SM_i) taking the form of arithmetic mean. Due to the fact that the groups of A, B, C and D normalization methods result in identical ordering (see Table 3), further analysis covered 10 methods of normalization: n1, n2, n3, n5, n5a, n8, n9, n9a, n11, n12a.

The article discusses the proposal of research procedure (section 3), based on Kendall's tau correlation coefficient and cluster analysis, which allows reducing the problem of normalization method selection for variable values.

The effects of simulation studies for 5 aggregation measures and 10 normalization methods were presented (section 4). Mean values of the adjusted Rand index taken from five models are in the interval $[0.845, 0.908]$. Therefore, the results of cluster analysis of normalization methods for the analysed aggregation measures are similar to each other (dendrite of cluster analysis results' similarity of normalization methods for five aggregation measures is presented in Figure 1).

The results of conducted research were illustrated by an empirical example presenting the application of five aggregation measures and ten normalization methods in linear ordering of Lower Silesian districts in terms of their tourism attractiveness level. Based on the analysis of the dendrograms three groups of normalization methods were separated. The biggest differences in the results of linear ordering refer to methods n2, n9a against the other normalization methods.

The author's own scripts, prepared in R environment, were applied in the calculations.

REFERENCES

BĄK, A., (1999). Modelowanie symulacyjne wybranych algorytmów wielowymiarowej analizy porównawczej w języku C++ [Simulation modeling of selected algorithms of multivariate comparative analysis with C++ language], Wrocław: Wydawnictwo Akademii Ekonomicznej we Wrocławiu, ISBN: 8370114016.

BORG, I., GROENEN, P. J. F., (2005). Modern multidimensional scaling, New York: Springer. ISBN: 978-0387-25150-9, http://dx.doi.org/10.1007/0-387-28981-X.

BORYS, T., (1984). Kategoria jakości w statystycznej analizie porównawczej [Category of quality in statistical comparative analysis], Prace Naukowe Akademii Ekonomicznej we Wrocławiu No. 284, Series: Monografie i opracowania No. 23. ISBN: 83-7011-000-0.

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

GENZ, A., AZZALINI, A., (2016). mnormt: The Multivariate Normal and t Distributions. *R package*, version 1.5-5, https://CRAN.R-project.org/package=mnormt.

GRABIŃSKI, T., (1984). Wielowymiarowa analiza porównawcza w badaniach dynamiki zjawisk ekonomicznych [Multivariate comparative analysis in research over the dynamics of economic phenomena], Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, Special series: Monografie No. 61, ISSN: 0209-1674.

GRABIŃSKI, T., (1992). Metody taksonometrii [Taxonometric methods], Kraków: Wydawnictwo Akademii Ekonomicznej w Krakowie.

GRABIŃSKI, T., WYDYMUS, S., ZELIAŚ, A., (1989). Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych [Numerical taxonomy methods in modeling socioeconomic phenomena], Warszawa: PWN, ISBN 83-208-0042-0.

GRYSZEL, P., WALESIAK, M., (2018). The application of selected multivariate statistical methods for the evaluation of tourism competitiveness of the Sudety communes, Argumenta Oeconomica, No. 1 (40), pp. 147–166, https://doi.org/10.15611/aoe.2018.1.06.

HELLWIG, Z., (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju i strukturę wykwalifikowanych kadr [Procedure of evaluating high level manpower data and typology of countries by means of the taxonomic method], Przegląd Statystyczny, Tom 15, z. 4, pp. 307–327.

HELLWIG, Z., (1972). Procedure of Evaluating High-Level Manpower Data and Typology of Countries by Means of the Taxonomic Method, [in:] Gostkowski Z. (ed.), Towards a system of Human Resources Indicators for Less Developed Countries, Papers Prepared for UNESCO Research Project, Ossolineum, The Polish Academy of Sciences Press, Wrocław, pp. 115–134.

HELLWIG, Z., (1976). Przechodniość relacji skorelowania zmiennych losowych i płynące stąd wnioski ekonometryczne [Transitivity of correlation and some econometric implications], Przegląd Statystyczny, Tom 23, z. 1, pp. 3–20.

HELLWIG, Z., (1981). Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych [Multivariate comparative analysis and applications in research of multifeature economic objects], In: W. Welfe (ed.), Metody i modele ekonomiczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną [Economic and mathematical methods and models in the improvement of socialist economy management], Warszawa: PWE, 46-68. ISBN 83-208-0042-0.

HUBERT, L., ARABIE, P., (1985). Comparing partitions, Journal of Classification, No. 1, pp. 193–218.

HWANG, C. L., YOON, K., (1981). Multiple attribute decision making – methods and applications. A state-of-the-art. Survey, New York: Springer-Verlag. ISBN: 978-3-540-10558-9, http://dx.doi.org/10.1007/978-3-642-48318-9.

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, pp. 105–112, Springer-Verlag, Berlin, Heidelberg, 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, https://dx.doi.org/10.1007/978-3-642-55721-7_12.

KENDALL, M. G., (1955). Rank correlation methods, London: Griffin.

KENDALL, M. G., BUCKLAND, W. R., (1986). Słownik terminów statystycznych [A dictionary of statistical terms], Warszawa: PWE, ISBN: 83-208-0504-X.

MILLIGAN, G. W., COOPER, M. C., (1988). A study of standardization of variables in cluster analysis, Journal of Classification, Vol. 5, No. 2, pp. 181–204.

NARDO, M., SAISANA, M., SALTELLI, A., TARANTOLA, S., HOFFMANN, A., GIOVANNINI, E., (2008). Handbook on Constructing Composite Indicators. Methodology and User Guide, Paris: OECD Publishing, ISBN: 978-92-64-04345-9.

PAWEŁEK, B., (2008). Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych [Normalization of variables methods in comparative research on complex economic phenomena], Kraków: Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, ISBN: 978-83-7252-398-3.

R CORE TEAM, (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria,

https://cran.r-project.org.

STEVENS, S. S., (1946). On the theory of scales of measurement, Science, Vol. 103, No. 2684, pp. 677–680.

WALESIAK, M., (1995). The analysis of factors influencing the choice of the methods in the statistical analysis of marketing data, Statistics in Transition, June, Vol. 2, No. 2, pp. 185–194.

WALESIAK, M., (2002). Uogólniona miara odległości w statystycznej analizie wielowymiarowej [The Generalized distance measure in multivariate statistical analysis], Wrocław: Wydawnictwo Akademii Ekonomicznej we Wrocławiu, ISBN: 83-7011-583-7.

WALESIAK, M., (2011). Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R [The Generalized distance measure GDM in multivariate statistical analysis with R], Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, ISBN: 978-83-7695-132-4.

WALESIAK M., (2014a). Wzmacnianie skali pomiaru w statystycznej analizie wielowymiarowej [Reinforcing measurement scale for ordinal data in multivariate statistical analysis], Taksonomia 22, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, No. 327, pp. 60–68.

WALESIAK M., (2014b). Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej [Data normalization in multivariate data analysis. An overview and properties], Przegląd Statystyczny, Tom 61, z. 4, pp. 363–372.

WALESIAK M., DUDEK A., (2018). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. *R package*, version 0.47-2, http://CRAN.R-project.org/package=clusterSim.

ZELIAŚ, A., (2002). Some notes on the selection of normalization of diagnostic variables, Statistics in Transition, Vol. 5, No. 5, pp. 787–802.