A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 286, 2013

Katarzyna Dębkowska *, *Marta Jarocka***

THE IMPACT OF THE METHODS OF THE DATA NORMALIZATION ON THE RESULT OF LINEAR ORDERING

Abstract. In taxonomy, various methods of the data normalization are used. They can significantly affect the result of the classification. This article presents the analysis of the impact of the methods of the data normalization on the result of linear ordering. The following methods and tools are used in this research: standardization, Weber's standardization, unitization, zero unitization with zero minimum and selected quotient transformation. A statement comparing the methods used in the study of classification is formulated. In the research data concerning innovation of European Union regions from Eurostat are used.

Key words: normalization, linear ordering.

I. INTRODUCTION

Normalization of the variables is a problem which is often considered in the literature in the field of multidimensional statistics. Most information about the normalization of the variables can be found in scientific papers containing the results of empirical comparative studies of complex economic phenomena that are preceded by a theoretical part describing the methodology used for supplying data for comparability. Detailed and in-depth discussion of normalization can be found in few works. The studies which are noteworthy in this field are K. Kukuła (2000) and B. Pawelek (2008).

This article presents the results of a research experiment, which aims to analyze the impact of various normalization procedures on the result of linear ordering of objects. The paper is based on normalization formulas including methods such as standardization, Weber's standardization, unitization, zero unitization and selected transformation quotients. Linear ordering was performed by means of model as well as non – model methods. Results are presented on the empirical example considering linear ordering on EU regions because of their innovative features.

* Ph.D., Chair of Business Informatics and Logistics, Bialystok University of Technology.
** Master, Chair of Business Informatics and Logistics, Bialystok University of Technology.

II. NORMALIZATION OF THE VARIABLES AS ONE OF THE STAGES OF LINEAR ORDERING

Normalization of the variables is one of the stages of linear ordering, and its task is to deprive titers features of their value and to unify the range of magnitude in order to bring them to comparability. The research experiment sought the answers to such questions as: how does changing normalization transformation formula affect the outcome of linear ordering, as well as which of the normalization methods give similar results of linear ordering.

The most commonly used normalization formulas were presented in Table 1. It also contains information on measurement scales of the input and output variables. The type of the scale of the input variables determines the set of acceptable normalization transformations (Jajuga, Walesiak, 2000), (Pawelek, 2008), (Walesiak, 2011).

Table 1. Types of variable normalization formulas

Sign.	Name	Formula	The scale of measurement variables	
			before normalization	after normalization
S	standardization	$z_{ij} = (x_{ij} - \bar{x}_j)/s_j$	ratio scale and (or) interval	interval
SW	Weber standardization	$z_{ij} = (x_{ij} - Me_j)/1.4826MAD_j$	ratio scale and (or) interval	interval
U	unitization	$z_{ij} = (x_{ij} - \bar{x}_j)/r_j$	ratio scale and (or) interval	interval
UZ	unitization with zero minimum	$z_{ij} = [x_{ij} - \min\{x_{ij}\})/r_j$	ratio scale and (or) interval	interval
QT max	quotient transformation	$z_{ij} = x_{ij} / \max_i\{x_{ij}\}$	ratio scale	ratio scale
QT s		$z_{ij} = x_{ij} / s_j$	ratio scale	ratio scale
QT r		$z_{ij} = x_{ij} / r_j$	ratio scale	ratio scale
QT $\bar{x}$		$z_{ij} = x_{ij} / \bar{x}_j$	ratio scale	ratio scale
QT sum		$z_{ij} = x_{ij} / \sum_{i=1}^{n} x_{ij}$	ratio scale	ratio scale
QT psk		$z_{ij} = x_{ij} / \sqrt{\sum_{i=1}^{n} x_{ij}^2}$	ratio scale	ratio scale

Source: M. Walesiak, *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wyd. UE we Wrocławiu, 2011, p. 19.

When deciding on normalization methods, besides the variables measurement scales, we should take into account the descriptive characteristics of the input diagnostic variables (Pawelek, 2008) and characteristics of the distribution of the variables, designated for normalized variable values (Walesiak 2011).

Normalization formulas such as unitization, zero unitization, ratio scale transformation with normalization basis which is equal to range are valuable because they provide normalized values of the variables varied variability (measured by standard deviation) and at the same time constant range for all variables.

Classic standardization, Weber's standardization and quotient transformation with the basis of normalization equal to standard deviation, cause unification of the variables in terms of variability which is measured by means of standard deviation (absolute median deviation) – this means the elimination of variation as a basis for differentiating objects.

Quotient transformations with the normalization basis equal to maximum and square root of the sum of squared observations provide a diverse: variability, arithmetic mean and range for normalized variable values.

Quotient transformations with the normalization basis equal to the sum and the arithmetic mean provide a diverse variability, range and constant arithmetic mean for normalized variables value.

All normalization formulas, being the linear observation transformations of each variable, keep the skewness and kurtosis of the variables' distribution and for each pair of variables all normalization formulas do not change the value of the Pearson's correlation coefficient.

III. EXPERIMENTAL RESEARCH RESULTS

The study involved 217 EU regions due to their innovative features. Preceded by a statistical analysis, the selection of diagnostic features was as follows:

– Personnel in the research and development sector as a percentage of total employment;

– EPO (European Patent Office) patent applications per million inhabitants;

– Expenditure on research and development per 1 inhabitant in the business sector (EURO per 1 inhabitant);

– Employment in science and technology as a percentage of the economically active population.

Realizations of diagnostic features were normalized according to different normalization formulas and next was performed a linear ordering of regions using non – model method by means of arithmetic mean. Table 2 shows the positions of the most innovative regions in the rankings obtained by using different normalization formulas.

Table 2. The positions of regions in the rankings
(linear ordering: non – model method by means of arithmetic mean)

Region's symbol	Signature of normalization formulas									
	S	**SW**	**U**	**UZ**	**QT max**	**QT s**	**QT r**	**QT $\bar{x}$**	**QT sum**	**QT psk**
DE11	1	1	1	1	1	1	1	1	1	1
DK01	2	3	2	2	2	2	2	2	2	2
DE21	3	2	3	3	3	3	3	3	3	3
SE11	4	4	4	4	4	4	4	4	4	4
DE14	5	5	5	5	5	5	5	5	5	11
DE71	6	8	8	8	8	6	8	10	10	12
DE91	7	9	6	6	6	7	6	6	6	6
DE12	8	7	7	7	7	8	7	7	7	16
FI18	9	13	9	9	12	9	9	14	14	18
FI1A	10	11	10	10	9	10	10	11	11	9

Source: own studies.

The data presented in Table 2 shows that the change in the normalization procedure affects the change of position in the analyzed region rankings. We can see them even in the top ten most innovative regions, for instance, regions marked by symbols DE71, DE12, FI18 changed their positions in the presented classifications even by a few places. In order to compare the rankings obtained in the research, Spearman correlation coefficients were calculated, whose values fluctuate in the range of 0.9624 –1.0000 (Table 3).

Table 3. Correlation matrix between result's rankings obtained by using different normalization formulas (linear ordering: non – model method by means of arithmetic mean)

Formula's sign.	**S**	**SW**	**U**	**UZ**	QT max	QT s	QT r	**QT $\bar{x}$**	QT sum	QT psk
S	1.000	0.9936	0.9997	0.9997	0.9986	1.0000	0.9997	0.9819	0.9819	0.9624
SW		1.0000	0.9946	0.9946	0.9961	0.9936	0.9946	0.9947	0.9947	0.9796
U			1.0000	1.0000	0.9994	0.9997	1.0000	0.9843	0.9843	0.9631
UZ				1.0000	0.9994	0.9997	1.0000	0.9843	0.9843	0.9631
QT max					1.0000	0.9986	0.9994	0.9886	0.9886	0.9669
QT s						1.0000	0.9997	0.9819	0.9819	0.9624
QT r							1.0000	0.9843	0.9843	0.9631
QT $\bar{x}$								1.0000	1.0000	0.9821
QT sum									1.0000	0.9821

Source: own studies.

Despite a high degree of correlation of linear ordering results, by means of different normalization methods we can observe significant differences in the results of particular rankings. Table 4 presents the number of regions whose positions have changed in relation to their position in the ranking obtained by means of standardization. These numbers correspond to the number of "shifts" of regions' positions in the analyzed rankings.

Table 4. Change in the positions of regions in the rankings
(linear ordering: non – model method by means of arithmetic mean)

Number of "shifts" position in the ranking of regions relative to ranking with standardization	Number of regions whose position has changed relative to ranking using standardization formula								
	SW	**U**	**UZ**	**QT max**	**QT s**	**QT r**	**QT $\bar{x}$**	**QT sum**	**QT psk**
0	16	69	69	32	217	69	8	8	15
1	44	87	87	62	0	87	18	18	18
2	36	43	43	48	0	43	23	23	17
3	18	8	8	25	0	8	15	15	16
4	13	8	8	13	0	8	16	16	13
5	16	1	1	14	0	1	16	16	13
6	12	0	0	9	0	0	21	21	9
7	11	0	0	6	0	0	12	12	13
8	10	0	0	3	0	0	7	7	8
9	5	0	0	2	0	0	6	6	6
10	7	2	2	1	0	2	7	7	3
…									
29	2	0	0	0	0	0	0	0	3
…									
83	0	0	0	0	0	0	0	0	1

Source: own studies.

The juxtaposition presented in Table 4 shows that with the change in the normalization procedure, positions of evaluated regions have changed as well. The biggest changes occurred in the case of replacement of the classic standardization by Weber's standardization or by selected quotient transformations. For example, a change of data normalization method from the standardization to Weber's standardization caused the fact that the position of two regions have shifted in the ranking list by 29 places, and when using quotient transformation QT psk the position of one of the regions changed up to 83 places.

Identical calculations were performed in the next stage of the research experiment, replacing the non – model linear ordering method by parametric Hellwig method (Hellwig, 1968). Tables 5, 6 and 7 present the results of the calculations which are equivalent to the results described in Tables 2, 3 and 4.

Table 5. The positions of regions in the rankings (linear ordering: parametric Hellwig method)

Region's symbol	Signature of normalization formulas									
	S	**SW**	**U**	**UZ**	**QT max**	**QT s**	**QT r**	**QT $\bar{x}$**	**QT sum**	**QT psk**
DE11	1	1	1	1	1	1	1	1	1	1
DE21	2	2	2	2	2	2	2	2	2	5
DK01	3	4	3	3	3	3	3	3	3	2
SE11	4	3	4	4	4	4	4	4	4	3
DE71	5	6	5	5	6	5	5	7	7	11
SE22	6	8	8	8	8	6	8	6	6	8
DE14	7	5	6	6	5	7	6	5	5	14
DE12	8	7	7	7	7	8	7	8	8	21
FI18	9	12	9	9	10	9	9	15	15	16
DE91	10	11	10	10	9	10	10	9	9	6

Source: own studies.

Table 6. Correlation matrix between result's rankings obtained by using different normalization formulas (linear ordering: parametric Hellwig method)

Formula's sign.	**S**	**SW**	**U**	**UZ**	QT max	QT s	QT r	**QT $\bar{x}$**	QT sum	QT psk
S	1.000	0.981	0.999	0.999	0.997	1.000	0.999	0.970	0.970	0.946
SW		1.000	0.983	0.983	0.987	0.981	0.983	0.996	0.996	0.968
U			1.000	1.000	0.999	0.999	1.000	0.973	0.973	0.947
UZ				1.000	0.999	0.999	1.000	0.973	0.973	0.947
QT max					1.000	0.997	0.999	0.981	0.981	0.955
QT s						1.000	0.999	0.970	0.970	0.946
QT r							1.000	0.973	0.973	0.947
QT $\bar{x}$								1.000	1.000	0.979
QT sum									1.000	0.979

Source: own studies.

Table 7. Change in the positions of regions in the rankings (linear ordering: parametric Hellwig method)

Number of "shifts" position in the ranking of regions relative to ranking with standardization	Number of regions whose position has changed relative to ranking using standardization formula								
	SW	**U**	**UZ**	**QT max**	**QT s**	**QT r**	**QT $\bar{x}$**	**QT sum**	**QT psk**
0	13	53	53	27	217	53	11	11	10
1	16	61	61	31	0	61	15	15	10
2	22	37	37	35	0	37	14	14	14
3	21	36	36	27	0	36	16	16	9
4	10	15	15	25	0	15	11	11	17
5	10	8	8	17	0	8	11	11	8
6	16	4	4	22	0	4	7	7	7
7	9	2	2	7	0	2	8	8	6
8	12	0	0	9	0	0	7	7	8
9	10	2	2	6	0	2	7	7	7
10	5	0	0	2	0	0	13	13	8
…									
35	2	0	0	0	0	0	1	1	2
…									
50	2	0	0	0	0	0	1	1	0
…									
60	0	0	0	0	0	0	1	1	1

Source: own studies.

Using the model linear ordering method leads to the same conclusions as in the case of application of non- model procedure. Changing the normalization method affects the positions of objects in the rankings, regardless of the classification procedure.

IV. SUMMARY

Summarizing the above considerations it must be explicitly stated that the choice of normalization formula influences the result of linear ordering, in both model and non – model methods. By the realization of a research experiment it was noted that, despite high levels of numerical correlation coefficients calculated between places in the rankings determined by using different methods of normalization, we can observe changes in the positions in the rankings in case of particular regions. Therefore, it is recommended to use the same normalization procedure of diagnostic variables while creating periodical classifications used to observe objects in time.

ACKNOWLEDGMENT
This scientific research was co-financed from the resources of the National Centre for Science nr N N111 530140.

REFERENCES

Hellwig Z. (1968), *Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr,* Przegląd Statystyczny, nr 4, pp. 307-326.

Kukuła K. (2000), *Metoda unitaryzacji zerowej*, Wydawnictwo Naukowe PWN, Warszawa.

Pawełek B. (2008), *Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych*, Wyd. Uniwersytetu Ekonomicznego w Krakowie.

Walesiak M. (2011), *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wyd. UE we Wrocławiu.

Jajuga K., Walesiak M. (2000), Standardization of Data Set under Different Measurement Scales [w:] Classification and Information Processing AT the Turn of the Millennium, red. R. Decker, W. Gaul, Springer Verlag, Berlin-Heidelberg.

Katarzyna Dębkowska, Marta Jarocka

WPŁYW FORMUŁY NORMALIZACYJNEJ NA WYNIK PORZĄDKOWANIA LINIOWEGO

W taksonomicznych badaniach empirycznych spotyka się różne formuły transformacji normalizacyjnej. Często pomija się fakt, iż jej wybór może istotnie wpływać na wynik klasyfikacji. W artykule zaprezentowano rezultaty eksperymentu badawczego, którego celem była analiza wpływu różnych procedur normalizacyjnych na wynik porządkowania liniowego obiektów. W pracy wykorzystano między innymi takie formuły normalizacyjne jak: standaryzacja, standaryzacja Webera, unitaryzacja, unitaryzacja zerowa oraz wybrane przekształcenia ilorazowe. Następnie dokonano komparacji otrzymanych wyników.

Procedurę badawczą oparto o dane zaczerpnięte z bazy Eurostat, które dotyczyły innowacyjności regionów Unii Europejskiej. Zjawisko innowacyjności regionów zostało scharakteryzowane poprzez szereg cech, których realizacje były dostępne na poziomie badanych jednostek terytorialnych.