CORE Metadata, citation and similar papers at core.ac.uk

ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 3(302), 2014

Artur Czech[*]

APPLICATION OF CHOSEN NORMALIZATION METHODS IN THE PROCESS OF CONSTRUCTION OF SYNTHETIC MEASURE IN INDIRECT CONSUMPTION RESEARCH

Abstract. The main purpose of this paper is to find the most effective method for the normalization of a final set of diagnostic variables for indirect consumption research with the use of synthetic consumption measures. The chosen normalization methods (standardization, unitarization and ratio transformation) were analyzed, both in the classical and the order approaches, with the use of different methods for multidimensional median vector construction (border and Weber median). Implementation of the multidimensional median construction is very important in the case of research objects with atypical characteristics of diagnostic variables (a separate analysis of Warsaw area in the Mazowieckie Voivodeship). This kind of research units can cause asymmetry in empirical distribution of diagnostic variables which has a significant impact on the result of the linear order set of research objects. Additionally, implementation of the Weber median allows for considering the interactions in a set of diagnostic variables, which is crucial from the point of view of economic analysis.

Keywords: normalization, indirect consumption, synthetic measure.

I. INTRODUCTORY REMARKS

There are two main research approaches to the analysis of consumption changes in literature (Słaby (2006a)):

- direct – analysis of the structure of material goods and services consumption based on the data gained from HBS (Households Budget Surveys) conducted by the Polish Central Statistical Office,
- indirect – analysis in which consumption is considered as indicator of social changes (e.g. living standard); researches are based on HBS data or data gained by other researchers.

The empirical assessments of indirect consumption (living standard) in Poland were conducted with the use of Geneva Distance Method, ELSI living standard measure (CBOS researches) and taxonomic synthetic measure.

[*] Ph.D., Department of Business Informatics and Logistics, Bialystok University of Technology.

Applications of taxonomic methods provide the most possibilities in the case of spatial diversification of consumption analysis. There are two ways of constructing synthetic measure with the use of both classic or order statistical measures. The former, was introduced by Hellwig (1968) and the latter is based on median which was first implemented by Lira, Wagner, Wysocki (2002). In the area of consumption research both methods were used by Słaby (2006b), Słaby and Czech (2011). It is very important to emphasize the fact that proposals of synthetic measure construction were both based on classical and order approach to standardization. Other approaches to normalization methods like unitarization, ratio transformation were not analyzed in the process of construction of indirect consumption measure.

The aim of the study is to produce verification of different normalization formulas with the use of multidimensional median vector in the conditions of skewness resulting from separate analysis of Warsaw area in the Mazowieckie Voivodeship indirect consumption analysis. It should be noticed that the need for separate analysis of Warsaw in Mazowieckie Voivodeship has already been presented in the literature by Słaby and Czech (2011).

II. THEORETICAL BASIS OF APPLIED NORMALIZATION METHODS

Particular diagnostic features which are used in synthetic measure construction usually have different denominations of their values as well as different range of magnitude. If classification methods, multidimensional calibration or linear ordering are employed, there is a need to bring different variables for comparison (Walesiak (2011)). This process is called normalization and can be introduced by applying classical and order statistical measures as the basis of transformation.

There are three basic types of normalization: standardization, unitarization and ratio transformation. The first one, in its classical version, uses arithmetic mean and standard deviation. The order version of standardization implements median and *mad* (median absolute deviation), where normalization takes the following form (Lira, Wagner, Wysocki (2002)):

$$z_{ij} = \frac{x_{ij} - \theta_j}{1{,}4826 \cdot mad(X_j)} \tag{1}$$

where:

$$mad\left(X_j\right) = \underset{i=1,2,\ldots,n}{med}\left|x_{ij} - \theta_j\right| \tag{2}$$

Values of θ_j are considered as particular elements of multidimensional median vectors (order and Weber). Weber median is estimated by using the following formulas:

$$T(\Theta, R^m) = \arg\min_{\Theta \in R^m} \left\{ \sum_{i=1}^{n} \left[\sum_{j=1}^{m} (x_{ij} - \theta_j)^2 \right]^{1/2} \right\} \tag{3}$$

It should be emphasized that there are other forms of construction of multidimensional median vector in the literature which can be applied in the process of normalization (Domański, Pruska, Wagner (1998)).

The second form of normalization is unitarization. In this kind of transformation its basis takes the form of range of variable. There are other, atypical unitarization proposals e.g. with the following formulas (Młodak (2006)):

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\max\limits_{i=1,2,\ldots,n} \left| x_{ij} - \bar{x}_j \right|} \tag{3}$$

In the case of order unitarization, the normalization is based on the formula (Młodak (2006)):

$$z_{ij} = \frac{x_{ij} - \theta_j}{\max\limits_{i=1,2,\ldots,n} \left| x_{ij} - \theta_j \right|} \tag{4}$$

The third form of normalization is called ratio transformation and it can be applied with the following formula:

$$z_{ij} = \frac{x_{ij}}{\bar{x}_j}. \tag{5}$$

On the other hand, in the case of order form of ratio transformation, arithmetic mean is replaced by median and normalization takes the following formulas:

$$z_{ij} = \frac{x_{ij}}{\theta_j} \tag{6}$$

All normalization formulas are considered linear transformations and should be implemented into variables which are measured on strong scales. It should be noticed that there are a lot of analyses of normalization formulas in literature with the application of R program, Walesiak (2011), Dębkowska, Jarocka (2013). But there is a lack of empirical analysis of implementation of multidimensional median vector in other forms than standardization.

III. PRESENTATION OF DIFFERENCES IN LINEAR ORDERING WITH IMPLEMENTATION OF DIFFERENT NORMALIZATION FORMULAS

The basis of synthetic measure construction with using different normalization formulas is a set of diagnostic variables drawn from Household Budget Surveys conducted by Polish Central Statistical Office. As the result of a variation and correlation analysis the final set of diagnostic variables was constructed and included: X_1 – the average monthly disposable income per capita in household (PLN), X_2 – share of healthcare expenditures in all expenditures (%), X_3 – share of transportation expenditures in all expenditures (%), X_4 – share of expenditures connected with occupying free time in all expenditures, (%), X_5 – the average threshold value of net monthly income considered as minimum (PLN). Artificial variable X_4 was created as aggregation expenditures on communication, recreation and culture, education, restaurants and hotels. Creation of this variable was essential due to the skewness of empirical distributions of the presented consumption expenditures and its influence on location measures for particular voivodeships.

Walesiak (2011) notices that during introduction of particular normalization formulas both measures of empirical distributions of particular variables before and after transformation should be taken into account. Basic statistic measures for final set of diagnostic variables are presented in Table 1.

Table. 1 Descriptive statistic measures of diagnostic variables

Variable	A_S	*Min*	*Max*	$R(X)$	$\bar{x}$	*m.b.*	*m.W.*	S_X	*mad*(*b*)	*mad*(*W*)
X_1	2.99	610.00	1367.55	757.55	807.64	785.56	795.87	157.35	38.89	36.78
X_2	–0.17	1.87	3.66	1.79	2.73	2.74	2.56	0.49	0.34	0.40
X_3	–0.51	2.30	6.74	4.45	4.85	5.03	4.74	1.19	0.84	0.78
X_4	2.39	12.67	20.22	7.54	14.52	14.13	14.36	1.72	0.64	0.84
X_5	1.59	1629.09	2416.17	787.08	1868.81	1825.48	1840.64	186.47	98.96	114.11

Explanations: A_S – skewness, *Min* – minimum, *Max* – maximum, $R(X)$ – range, $\bar{x}$ – arithmetic mean, *m.b.* – border median, *m.W.* – Weber median, S_X – standard deviation, *mad(b)* – median absolute deviation (border median), *mad*(*W*) – median absolute deviation (Weber median).

Source: own studies.

The data presented in the table shows that the variables X_1, X_4, X_5 have strong skewness that influences the value of arithmetic mean. Implementation of border median decreases values of location statistics. Additionally, application of Weber median allows for interactions in a set of diagnostic variables which in turn increases values of location statistics in relation to border median. The process of standardization, unitarization and ratio transformation both in classic and order form received sets of normalized variables. Location and variation statistics of normalized variables are presented in Table 2.

Table 2. Statistical measures for normalized variables

Type of normalization		Arithmetic mean/median					S_Z / *mad*(*Z*)				
		Z_1	Z_2	Z_3	Z_4	Z_5	Z_1	Z_2	Z_3	Z_4	Z_5
Classical	*S*	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
	U	0.00	0.00	0.00	0.00	0.00	0.28	0.53	0.47	0.30	0.34
	RT	1.00	1.00	1.00	1.00	1.00	0.20	0.18	0.25	0.12	0.10
Order (border median)	*S*	0.00	0.00	0.00	0.00	0.00	0.67	0.67	0.67	0.67	0.67
	U	0.00	0.00	0.00	0.00	0.00	0.07	0.37	0.31	0.11	0.17
	RT	1.00	1.00	1.00	1.00	1.00	0.05	0.12	0.17	0.05	0.05
Order (Weber median)	*S*	–0.19	0.30	0.26	–0.18	–0.09	0.71	0.57	0.73	0.52	0.59
	U	–0.02	0.16	0.12	–0.04	–0.03	0.07	0.31	0.35	0.11	0.17
	RT	0.99	1.07	1.06	0.98	0.99	0.05	0.13	0.18	0.05	0.05

Explanations: *S* – standardization, *U* – unitarization, *RT* – ratio transformation, S_Z – standard deviation, *mad*(*Z*) – median absolute deviation.

Source: own studies.

Implementation of order normalization formulas shows that only standardization with border median results in unification of the variables in terms of variability measured by means of median absolute deviation. It means the elimination of variation as a basis for differentiating analyzed objects. In the case of order normalization methods with Weber median does not bring median of normalized variable to zero and *mad* does not equal one. It means that this kind of normalization does not produce unification of variation measured by means of median absolute deviation. Discussion about deviation of normalized variables with the use of Weber median from basic standardization assumptions is presented by Lira, Wagner, Wysocki (2002):

Implementation of presented formulas in the process of normalization of final set of diagnostic variables in the process of living standard synthetic measure construction allowed for creation of ranking. The locations of particular voivodeships are presented in Table 3.

Table 3. The positions of voivodeship in the ranking

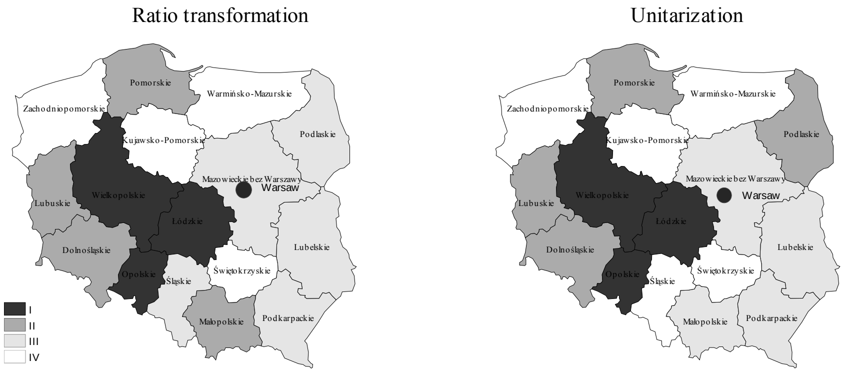
Research objects	*MK*(.)			*MP*1(.)			*MP*2(.)		
	S	*U*	*RT*	*S*	*U*	*RT*	*S*	*U*	*RT*
dolnośląskie	6	7	7	4	7	7	5	7	7
kujawsko-pomorskie	13	14	14	13	14	14	12	14	14
lubelskie	10	10	11	12	10	11	13	10	11
lubuskie	8	8	6	6	8	6	6	8	6
łódzkie	2	3	4	2	3	4	2	3	4
małopolskie	9	9	8	11	9	8	11	9	8
mazowieckie without Warsaw	11	11	10	10	11	10	10	11	10
opolskie	3	1	2	5	1	2	4	1	2
podkarpackie	14	13	13	17	13	13	17	12	13
podlaskie	7	6	9	8	6	9	9	6	9
pomorskie	5	5	5	3	5	5	3	5	5
śląskie	12	12	12	9	12	12	8	13	12
świętokrzyskie	15	15	15	16	15	15	16	15	15
warmińsko-mazurskie	17	17	17	15	17	17	15	17	17
wielkopolskie	4	2	3	7	2	3	7	2	3
zachodniopomorskie	16	16	16	14	16	16	14	16	16
Warsaw	1	4	1	1	4	1	1	4	1

Explanations: *MK*(.) – classical synthetic measure, *MP*1(.) – order synthetic measure with border median, *MP*2 – order synthetic measure with Weber median, *S* – standardization, *U* – unitarization, *RT* – ratio transformation.

Source: own studies.

It is to be noticed that the ranking made by using of unitarization (equations 3–4) locates Warsaw as forth. This shows that this kind of normalization is not proper in linear ordering because Warsaw is considered as one of the most developed cites in Europe in the area of consumption. Synthetic measures can be implemented in construction of similar areas of living standard. The example of this kind of classification is presented in the picture 1. The division of research objects into four groups was made by applying of the three median formula (Młodak 2006). In order to compare distributions of synthetic measures, Pearsons correlation coefficients were calculated (Table 4).

The Spearman correlation coefficients have been calculated as well. Strong correlation of both classical and order synthetic measures in two types of the correlation analysis can be considered as the lack of significant statistical differences in the values of synthetic measures.



Picture 1. Spatial diversification of indirect consumption with Weber median
Source: own studies.

Table 4. Matrix of Pearson's coefficients and results of Friedman test

Pearson's correlation coefficients										Chi square. ANOVA (N = 17, df 8) =81,06667; p = 00000; F it coefficient = ,59608; rank mean = 0,57083			
	MK(S)	MK(U)	MK(RT)	MP1(S)	MP1(U)	MP1(S)	MP2(S)	MP2(U)	MP2(RT)	mean rank	sum of rank	arithmetic mean	standard deviation
MK(S)	1.00	0.93	0.98	0.85	0.93	0.98	0.86	0.94	0.97	6.00	102	0.23	0.11
MK(U)	0.93	1.00	0.95	0.62	1.00	0.94	0.62	1.00	0.95	7.59	129	0.26	0.13
MK(RT)	0.98	0.95	1.00	0.80	0.94	1.00	0.81	0.96	1.00	8.47	144	0.29	0.14
MP1(S)	0.85	0.62	0.80	1.00	0.61	0.82	1.00	0.64	0.79	2.24	38	0.16	0.15
MP1(U)	0.93	1.00	0.94	0.61	1.00	0.94	0.62	1.00	0.95	3.76	64	0.18	0.15
MP1(RT)	0.98	0.94	1.00	0.82	0.94	1.00	0.83	0.96	1.00	4.94	84	0.21	0.16
MP2(S)	0.86	0.62	0.81	1.00	0.62	0.83	1.00	0.65	0.80	3.65	62	0.17	0.15
MP2(U)	0.94	1.00	0.96	0.64	1.00	0.96	0.65	1.00	0.97	2.82	48	0.17	0.15
MP2(RT)	0.97	0.95	1.00	0.79	0.95	1.00	0.80	0.97	1.00	5.53	94	0.22	0.16

Source: own studies.

In order to exclude the suspicion about equal statistical distributions of synthetic measures, the random samples Friedman test was used. This kind of non-parametrical test is considered as extension of Wilcoxon rank test. A detailed presentation of it can be found in Aczel (2000) and Stanisz (2006). The results of this test are presented in Table 5. There is very low critical level of significance in the process of verification of null hypothesis that particular distributions of synthetic indirect consumption measures are the same. This resulted in rejection of null hypothesis and it can be noticed that particular distributions of synthetic measures are different.

IV. CONCLUSIONS

The use of different normalization methods both in classical and order form influences the result of linear ordering. In the research process was noted that introducing proper normalization formulas should be preceded by research of skewness of empirical distributions of diagnostic variables. Standardization with border median causes unification of all variables by means of variability measured by median absolute deviation. Implementation normalization formulas based on border median can cause decreasing of information and research value of constructed models. In the conditions of strong skewness caused by untypical values of diagnostic variables implementation of unitarization based on maximum absolute deviation can cause no proper linear ordering.

REFERENCES

Aczel D. A. (2000), *Statystyka w zarządzaniu*, PWN, Warszawa, p. 737–742.

Domański Cz., Pruska K., Wagner W. (1998), *Wnioskowanie statystyczne przy nieklasycznych założeniach*, Wyd. Uniwersytetu Łódzkiego, Łódź, p. 181–186.

Dębkowska K., Jarocka M. (2013), The impact of the methods of the data normalization on the result of linear ordering, Methods and applications of multivariate statistical analysis, Acta Universitatis Lodziensis, Folia Oeconomica 286, p. 181–188.

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, p. 307–326.

Lira J., Wagner W., Wysocki F. (2002), Mediana w zagadnieniach porządkowania obiektów wielocechowych, *Statystyka regionalna w służbie samorządu lokalnego i biznesu*, Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej, Akademia Ekonomiczna, Wrocław, p. 87–99.

Młodak A. (2006), *Analiza taksonomiczna w statystyce regionalnej*, Difin, Warszawa, p. 36–42.

Słaby T. (2006a), Statystyczny pomiar konsumpcji, in: M. Janoś-Kresło i B. Mróz (edit.), *Konsument i konsumpcja we współczesnej gospodarce,* Wyd. SGH, Warszawa, p. 81.

Słaby T. (2006b), *Konsumpcja, Eseje statystyczne,* Difin, Warszawa, p. 117–131.

Słaby T., Czech A. (2011), Zróżnicowanie regionalne konsumpcji w ujęciu pośrednim – ujęcie statyczne i przestrzenno-czasowe, *Studia i Prace Kolegium Zarządzania i Finansów*, SGH, Warszawa, p. 7–22.

Stanisz A. (2006), *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny*, Tom 1. Statystyki podstawowe, StatSoft, Kraków, p. 392–398.

Walesiak M. (2011), *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R,* Wyd. UE, Wrocław, p. 18–21.

Artur Czech

ZASTOSOWANIE WYBRANYCH FORMUŁ NORMALIZACYJNYCH W PROCESIE BUDOWY MIERNIKA SYNTETYCZNEGO W BADANIACH KONSUMPCJI W UJĘCIU POŚREDNIM

Praca ma na celu poszukiwanie najwłaściwszej metody normalizacji finalnego zestawu cech diagnostycznych do badań konsumpcji w ujęciu pośrednim z użyciem ocen syntetycznych. Analizie poddano wybrane formuły normalizacyjne (standaryzacja, unitaryzacja, przekształcenie ilorazowe) zarówno w ujęciu klasycznym i pozycyjnym z zastosowaniem różnych sposobów konstrukcji wielowymiarowego wektora medianowego (mediana brzegowa i Webera). Wykorzystanie pojęcia mediany wielowymiarowej jest bardzo istotne w przypadku uwzględnienia w analizie przestrzennej obiektów badania o nietypowych wartościach cech diagnostycznych (postulat oddzielnego traktowania Warszawy w analizach w ujęciu województw). Istnienie tego typu jednostek powoduje występowanie asymetrii rozkładu, co ma bardzo istotny wpływ na wynik porządkowania liniowego zbioru obiektów podanych analizie. Zastosowanie mediany Webera dodatkowo pozwala na uwzględnienie interakcji w zbiorze zmiennych diagnostycznych, co jest niezwykle istotne z punktu widzenia prowadzonych analiz ekonomicznych.