LUBELSKI ROCZNIK PEDAGOGICZNY
T. XLI, z. 4 – 2022

DOI: 10.17951/lrp.2022.41.4.91-101

Sławomir Pasikowski
University of Łódź
ORCID – 0000-0002-0768-1596

NORMALIZATION TRANSFORMATIONS OF DATA IN THE PROCEDURE OF STATISTICAL METHODS USED IN EDUCATION RESEARCH*

Introduction: Sometimes the mapping of the variable values made during the measurement needs to be transformed due to the conditions related to the data analysis.
Research Aim: The aim of this article is to present the issue of normalization transformations and the circumstances of their application.
Evidence-based Facts: Due to lack of knowledge or experience, transformations are not taken into account in data analysis in education research. And if they do occur, among the most popular methods of data transformation, are chosen those which reduce the skewness of distributions. However, the method of transformation is not always chosen adequately to the properties of the data and the conditions of analysis. Normalization transformations are among the simpler yet effective solutions for preparing data for analysis.
Summary: Normalization transformations minimize the risk of artifacts due to differences in orders of magnitude and units of measurement. This is particularly important when conducting analyses using multidimensional scaling and multivariate classification methods.

Keywords: data transformations, normalization, statistical methods.

INTRODUCTION

The empirical data collected is a representation of the values of the variable or variables that have been observed. The quality of this representation depends mainly on such factors as sampling error, method error, measurement error, or measurement scale. The last of these factors is responsible for the form and quality of the representation of variable values in the set of symbols that are their

* Suggested citation: Pasikowski, S. (2022). Normalization Transformations of Data in the Procedure of Statistical Methods Used in Education Research. *Lubelski Rocznik Pedagogiczny*, *41*(4), 91–101. http://dx.doi.org/10.17951/lrp.2022.41.4.91-101

representations. Variable values can be carried out in a set of numbers, or a set of symbols other than numbers. This process is described in detail by the mathematical theory of functions and the theory of category. Sometimes the mapping of variable values made during measurement requires transformation due to conditions related to data analysis (cf. Tabachnick and Fiedel, 2013; Tinsley and Brown, 2000; Venter and Maxwell, 2000). Transformation of a dataset is a map of this set onto itself, but in the other system of symbols, and the conditions mentioned include meeting the assumptions of analysis methods in terms of the properties of the distribution of the variable, unifying orders of magnitude and units of measurement, linearization of non-linear models. Without consideration of these conditions, erroneous results of the analysis are obtained, but it also happens that the conduct of the analysis becomes seriously complicated due to the limited comparability of the variable distributions under study. Among the simplest transformations is translation, which involves either adding or subtracting a constant value from each observation in the data set so that a tidy reference point, such as the arbitrary point 0, is obtained. However, the transformations used in data analysis in social research are associated with misunderstandings over the issue of the form of transformations and the circumstances of their application. Therefore, the purpose of this article is to provide an overview of this issue. Due to the breadth of the topic, the focus of this article will be on normalization transformations (in a broad sense), since they are the ones that usually secure the possibility of conducting multivariate analyses by meeting the condition of comparability of variable distributions, but also facilitate the performance of operations on data and the interpretation of measurement results.

PURPOSE AND CONDITIONS OF NORMALIZATION TRANSFORMATIONS

The purpose of normalization transformations is to unify orders of magnitude and units of measurement. This makes it possible to compare distributions of variables expressed in different units of measurement or different orders of magnitude. The limiting condition for the use of normalization is that the power of the scale of measurement of the original variable must not be less than the power of the interval scale. The exception is rank normalization, which can be applied to data measured on an ordinal scale due to the fact that the rank itself transforms data from measurement on an ordinal scale to data corresponding to an interval scale and allows independence from the distribution of the variable. The reason for the aforementioned limitation is that there is no justification for conducting normalizing transformations when the scale of measurement does not provide for the possibility of determining equality of intervals, equality of differences and equality

of ratio of the measured variable values. This is the case when, for example, formal education is measured as a variable taking values ordered by "greater than" or "less than" relations, but not equality of differences and intervals. For the latter, a fixed unit of measurement is necessary, and such is lacking when the operationalization of this trait provides for such values as primary, secondary, tertiary. The situation is changed by the introduction of either one year or one semester as the unit of measurement of formal education. Then, with the accuracy of the adopted unit of measurement, it is possible to determine the ranges of values and the length of the interval taken by the difference between the values. Moreover, this way of operationalizing this variable determines the point 0, which is natural in the sense that the number 0 means that formal education measured in units of time takes on a value that corresponds to an empty set of years. Put simply, it takes on the value "no years". However, if the purpose of measuring education measured on an ordinal scale was to compare the distributions of two or more populations that differ in the number of values that formal education can take, or it would be reasonable to assume differences in the length of the stages of educational attainment, then rank normalization should be used. A didactic example of such normalization, taking tied ranks into account at the same time, is presented in Table 1.

Without normalizing the data, parametric characterization of the FE3 and FE4 distributions would have to be limited to positional statistics, and their parametric comparison would give a distorted result. This becomes particularly apparent when relating the results: no differences in the range of variability (q, v_q), and markedly different as_{Yule} values, to the sequence of observations in the x_{FE3} and x_{FE4} columns. While, in fact, both distributions are right-skewed and have similar but, nevertheless, different variability.

Table 1.
Rank normalization of source data and structure description of their distributions

Sort order	x_{FE3}	x_{FE4}	$rank_{FE3}$	$rank_{FE4}$
1	1	1	2.5	2
2	1	1	2.5	2
3	1	1	2.5	2
4	1	2	2.5	5
5	2	2	6	5
6	2	2	6	5
7	2	3	6	7.5
8	3	3	9	7.5
9	3	4	9	9.5
10	3	4	9	9.5

sum	.	.	55	55
m	.	.	5.50	5.50
me	2	2	6	5
q_1	1	1.25	2.5	2.75
q_3	2.75	3	8.25	7.5
d_1	1	1	2.5	2
d_9	3	4	9	9.5
s	.	.	2.86	2.93
q	0.875	0.875	2.875	2.375
v	.	.	0.520	0.534
v_q	0.438	0.438	0.479	0.475
as	.	.	0.12	0.08
as_{Yule}	-0.14	0.14	-0.22	0.05
k	.	.	-1.81	-1.41
k_{dq}	1.14	1.71	1.13	1.58

FE3 – variable "formal education" with three values, FE4 – variable "formal education" with four values, sum – sum, m – arithmetic mean, me – median, q_1 – first quartile, q_3 – third quartile, s – standard deviation, q – quartile deviation, d_1 – first decile (10th percentile), d_9 – ninth decile (90th percentile), v – coefficient of variation, v_q – quartile coefficient of variation, as – classical asymmetry coefficient, as_{Yule} – Yule's positional asymmetry coefficient calculated on basis of quartiles, k – classical kurtosis coefficient, k_{dq} – positional kurtosis coefficient calculated on the basis of the first and ninth deciles and the first and third quartiles

Source: Author's own study.

However, most normalization transformations are applicable to measurement scales of higher power (ordinal, ratio). Because they share common properties, they can be expressed in a single model.

NORMALIZATION TRANSFORMATIONS MODEL

The general model of normalization transformations takes the following form (Jajuga and Walesiak, 2000; Walesiak, 2014):

$$z_{ij} = b_j x_{ij} + a_j$$

where
$a = -\mu_j/\sigma_j$
$b = 1/\sigma_j$, which can be converted to a simpler form (cf. *ibid.*):

$$z_{ij} = \frac{1}{\sigma_j} x_{ij} + \left(\frac{-\mu_j}{\sigma_j}\right) = \frac{x_{ij} - \mu_j}{\sigma_j}$$

z_{ij} – transformed value of the j-th variable for the i-th object
x_{ij} – value of the j-th variable for the i-th object
μ_j – shift parameter to the contractual zero for the j-th variable
σ_j – scale parameter for the j-th variable ($\sigma_j > 0$)

The parameter μ allows the location of observations relative to a contractual zero point. The estimator of this parameter can be, for example, the arithmetic mean or median. The parameter σ determines the range within which this localization occurs. The estimator of this parameter can be, for example, the standard deviation or the range. In this case, it is proper to associate two properties of the variable distribution: central tendency and variability.

In the case of quotient transformations, the parameter μ is 0. This is understandable when considering that they transform variables measured on a ratio scale to a variable measured on the same scale. The ratio scale has a natural zero, and this means that the distance of an observation from a point of natural zero is informed by the sum of the units of measure corresponding to that observation, i.e. its value. Table 2 presents the normalization transformation formulas that are most commonly encountered in the literature (cf. Jajuga and Walesiak, 2000; Walesiak, 2012, 2014).

Formulas 1 through 11 apply to data from measurements on an interval scale and a quotient scale. After normalization with these formulas, the scale of measurement is interval. Formulas 12 through 19 apply to data derived from measurements on the quotient scale. They transform this data into a form which measurement scale is also quotient. Thus, it is understood that the scale of measurement at the input and at the output is the main criterion for selecting a normalization formula.

Subsequent criteria make the choice of the formula dependent on expectations of the distributions characteristics after the transformation, as well as the location of the contractual zero, such as at the level of average value: arithmetic mean or median (transformations 1, 2, 3, 4, 5, 7, 8, 10), the mid of range (transformation 9), or minimal value (transformation 6). When the purpose of the transformation is to stabilize the mean (*m*, *me*) of the transformed variables, but while maintaining variation in variability and range, formulas 8, 10, 11, 13, 14, 18 become helpful. On the other hand, when the goal is to stabilize and standardize variability, transformations 1, 2, 7, 15, 17 give fit results. Varying variability, but a constant range for all variables provide transformations 4, 5, 6, 16. Varying variability, but with it varying arithmetic mean and range provide transformations 12 and 19.

If the purpose of normalization is primarily to standardize the order of magnitude justified, for example, by the convenience of comparisons and compilations,

then formulas 1, 2, 6, 12, 16, as well as transformations using the natural logarithm and decimal logarithm, work well in this regard. The referenced criteria gain confirmation in simulations using Excel.

Table 2.
Source data transformation formulas

No.	Transformation	Formula*
1	Standardization [z(x)]	$z = (x - m)/s$
2	Positional standardization [$z_q(x)$]	$z = (x - me)/mad$
3	Weber's standardization [$z_{Weber}(x)$]	$z = (x - me)/(1.4826)mad$
4	Unitization [u(x)]	$z = (x - m)/r$
5	Positional unitization [$u_q(x)$]	$z = (x - me)/r$
6	Reset unitization [$u_r(x)$]	$z = (x - min_x)/r$
7	Normalization [nr(x)]	$z = (x - m)/\sqrt{\Sigma(x-m)^2}$
8	Positional normalization [$nr_q(x)$]	$z = (x - me)/\sqrt{\Sigma(x-me)^2}$
9	Normalization with 0 centrally located [$nr_0(x)$]	$z = (x - mr)/(r/2)$
10	Normalization in the range [$nr_{(-1,1)}(x)$]	$z = (x - m)/max_{\|x-m\|}$
11	Positional normalization in the range [$nr_{q(-1,1)}(x)$]	$z = (x - me)/max_{\|x-me\|}$
12	Quotient 1	$z = x/max_x$
13	Quotient 2	$z = x/m$
14	Quotient 3	$z = x/me$
15	Quotient 4	$z = x/s$
16	Quotient 5	$z = x/r$
17	Quotient 6	$z = x/mad$
18	Quotient 7	$z = x/\Sigma x$
19	Quotient 8	$z = x/\sqrt{\Sigma x^2} = x/(\Sigma x^2)^{1/2}$

x – value of variable in the data set (observation), z – transformed x, m – arithmetic mean, me – median, min_x – minimal x in the data set, max_x – maximal x in the data set, s – standard deviation, r – range (difference between max_x and min_x), mr – mid-range calculated as: (max + min)/2, $max_{|x-m|}$ – maximal absolute difference between observations and arithmetic mean in the data set, mad ($me_{|x-me|}$) – median absolute deviation calculated as a median of absolute differences between observations and median in the data set, Σ – sum
*the formulas are given in a linear notation.

Source: Author's own study based on (Walesiak, 2012, 2014).

SELECTED PROPERTIES OF NORMALIZATION TRANSFORMATION MODELS

The two basic properties of normalization formulas (Walesiak, 2014) are (1) they do not change symmetry and kurtosis of distribution, and (2) for each pair of variables, they do not change the value of the linear correlation. The justification for these properties can be formulated based on the analysis of the general normalization transformation model. However, empirical confirmation already requires experimentation. For this purpose, a set of randomly generated data was used. It consisted of 50 records, each of which contained two values. In other words, the set contained 50 observations corresponding to one variable (X) and 50 observations corresponding to the other variable (Y). In the first step, the form of the distribution of the one-dimensional random variable X was observed. In the second, Pearson's *r* statistic used in the parametric characterization of the distribution of the two-dimensional random variable XY was observed. The results are in agreement with the two basic properties reported in the literature.

Whilst testing property (1), the symmetry and kurtosis of the distributions after transforming the original data with each of the 19 formulas did not change (respectively: *as* = -0.01, *k* = -0.91), although the distributions differ in central tendency and variability (Table 3).

Table 3.
Descriptive statistics of source data and its transformations

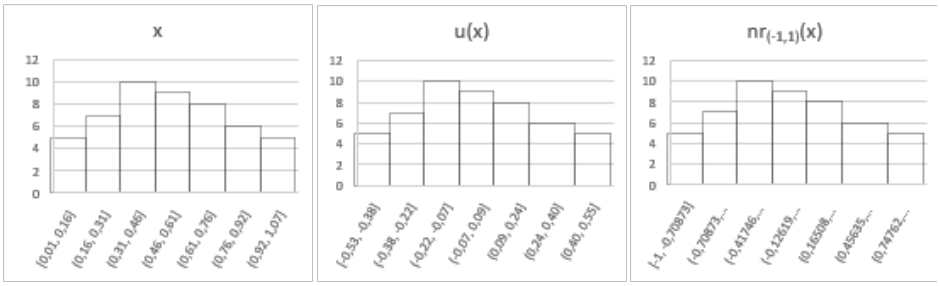
Distribution	*m*	*s*	*as*	*k*
X	0.53	0.27	–0.01	–0.91
1	0.00	1.00	–0.01	–0.91
2	–0.10	1.30	–0.01	–0.91
3	0.00	0.04	–0.01	–0.91
4	0.00	0.27	–0.01	–0.91
5	–0.02	0.27	–0.01	–0.91
6	0.53	0.27	–0.01	–0.91
7	0.00	0.14	–0.01	–0.91
8	–0.01	0.11	–0.01	–0.91
9	0.06	0.53	–0.01	–0.91
10	0.00	0.52	–0.01	–0.91
11	–0.04	0.50	–0.01	–0.91
12	0.53	0.27	–0.01	–0.91

Distribution	*m*	*s*	*as*	*k*
13	1.00	0.51	–0.01	–0.91
14	0.96	0.49	–0.01	–0.91
15	1.96	1.00	–0.01	–0.91
16	0.54	0.27	–0.01	–0.91
17	2.54	1.30	–0.01	–0.91
18	0.02	0.01	–0.01	–0.91
19	0.13	0.06	–0.01	–0.91

Source: Author's own study.

Selected figures (chosen due to space limitations) visualize the shape of the distributions (Figure 1). The distributions are presented as a histogram for continuous (interval) series. The number of classes (*k*) was calculated according to the standard formula: k = 1 + 3,322 log(n), where *n* is sample size. Similarly, according to the standard formula, the width of classes (*h*) was calculated: h = (max–min)/k.

Figure 1.
Distribution of source data (x) and transformed



Source: Author's own study.

Whilst testing property (2), the results showed that the correlation coefficient also did not change (Pearson's $r = 0.45$). Moreover, it can be noted that the correlation coefficient between the distributions obtained by applying different transformation formulas was the same as for the distributions of the original two-dimensional variable XY and its transformations according to the same formula (Table 4).

Table 4.
Linear correlation coefficient Pearson's r of source variable and its transformations

V	X	1_X	2_X	3_X	4_X	5_X	6_X	7_X	8_X	9_X	10_X	11_X	12_X	13_X	14_X	15_X	16_X	17_X	18_X	19_X
Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
1_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
2_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
3_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
4_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
5_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
6_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
7_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
8_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
9_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
10_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
11_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
12_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
13_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
14_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
15_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
16_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
17_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
18_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
19_Y	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45

V – variable, X – variable X, Y – variable Y, 1_X to 19_X – numbers of subsequent transformations of X variable, 1_Y to 19_Y – numbers of subsequent transformations of Y variable

Source: Author's own study.

Symmetry, flatness and co-variability of distributions are not altered by the presented normalization formulas. This may be an advantage, but if there is a need to change these properties, then other transformations must be used, such as logarithmic, exponential, root, power and reciprocal, and in the case of weaker measurement scales – logit or angular transformations. Nevertheless, some of the presented formulas stabilize the variance of the distribution, while others bring the value of the variance measures to the same order of magnitude. These can be useful when the analysis methods require meeting the assumption of homoscedasticity.

SUMMARY AND CONCLUSIONS

This article discusses the normalization transformation by variables. However, the same normalization method can be carried out by object. Then μ is the shift to the conventional zero for the *i*-th object, and σ is the scale parameter for the *i*-th object. In pedagogy and related disciplines, it is less common to find such normalization, although it can certainly be very useful in research conducted in a single-case design.

The above transformations can be performed in Excel, SPSS, Statistica and many other programs by entering in the command lines the formulas given above. Also, there can be used the *clusterSim* package in *R* using the *data.Normalization* function. However, the normalizing transformation has basic limitation which is problematic interpretation of transformed data in the language of original data. Under certain conditions, this can be remedied by transforming the transformation formula. Then in place of z_{ij} getting a value expressed in units of the source data. An example of transforming the formula z = (x m)/s according to the scheme: [1] (z = (x m)/s)s, [2] zs = x m, [3] zs + m = x. Another limitation, usually very serious, is the non-unimodal distribution of variable before transformation. The inability to transform data from a weaker power scale to a higher power scale can also be considered a limitation, which means that in research projects that take into account multiple variables, the transformation results in a reduction of the power of measurement scale to the level of the variable with the weakest power. Despite these limitations, normalization transformations make it possible to rationally perform analyses minimizing the risk of artifacts due to differences in orders of magnitude and units of measurement. This is particularly important for multivariate classification methods and multidimensional scaling, and wherever there is a need to compare distributions.

REFERENCES

Jajuga, K., Walesiak, M. (2000). Standardisation of Data Set under Different Measurement Scales. In G.W. Decker (Ed.), *Classification and Information Processing at the Turn of the Millennium* (pp. 105–112). Springer-Verlag.

Tabachnick, B.G., Fiedel, L.S. (2013). *Using Multivariate Statistics*. Pearson.

Tinsley, H.E., Brown, S.D. (2000). Multivariate Statistics and Mathematical Modeling. In H.E. Tinsley, S.D. Brown (Eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 3–36). Academic Press.

Venter, A., Maxwell, S.E. (2000). Issues in the Use and Application of Multiple Regression Analysis. In H.E. Tinsley, S.D. Brown (Eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 151–182). Academic Press.

Walesiak, M. (2012). Podstawowe własności analizy wielowymiarowej. In M. Walesiak, E. Gatnar (Eds.), *Statystyczna analiza danych z wykorzystaniem programu R* (pp. 62–80). PWN.

Walesiak, M. (2014). Przegląd formuł normalizacji wartości zmiennych oraz ich oraz ich własności w statystycznej analizie wielowymiarowej. *Przegląd statystyczny*, *61*(4), 363–372.

PRZEKSZTAŁCENIA NORMALIZACYJNE DANYCH W PROCEDOWANIU METOD STATYSTYCZNYCH WYKORZYSTYWANYCH W BADANIACH NAD EDUKACJĄ

Wprowadzenie: Zdarza się, że odwzorowanie wartości cechy dokonane podczas pomiaru wymaga przekształcenia z uwagi na warunki związane z analizą danych.
Cel badań: Celem artykułu jest przybliżenie zagadnienia normalizacyjnych przekształceń danych i okoliczności ich zastosowań.
Stan wiedzy: Z powodu braku wiedzy lub doświadczenia przekształcenia danych mogą nie być brane pod uwagę podczas analiz prowadzonych w badaniach nad edukacją. Jeśli występują, to spośród najpopularniejszych sposobów przekształceń stosunkowo często wybierane są te, które służą redukcji skośności rozkładów. Jednak nie zawsze sposób przekształcenia dobierany jest adekwatnie do własności danych oraz warunków analizy. Przekształcenia normalizacyjne należą do prostszych, a zarazem efektywnych rozwiązań przygotowujących dane do prowadzenia analiz.
Podsumowanie: Przekształcenia normalizacyjne minimalizują ryzyko powstawania artefaktów na skutek różnic w zakresie rzędu wielkości oraz jednostek miary. Ma to szczególne znaczenie podczas prowadzenia analizy z użyciem wielowymiarowego skalowania i wielowymiarowych metod klasyfikacji.

Słowa kluczowe: przekształcenia danych źródłowych, normalizacja, metody statystyczne.