STANDARDISATION OF DATA SET UNDER DIFFERENT MEASUREMENT SCALES

Krzysztof Jajuga[1], Marek Walesiak[1]

[1] Wrocław University of Economics,
Komandorska 118/120, 53-345 Wrocław, Poland

Abstract: Standardisation of multivariate observations is the important stage that precedes the determination of distances (dissimilarities) in clustering and multidimensional scaling. Different studies (e.g. Milligan and Cooper (1988)) show the effect of standardisation on the retaining of cluster structure in various data configurations. In the paper the survey of standardisation formulas is given. Then we consider the problem of different scales of measurement and their impact on:

– the selection of the standardisation formula;

– the selections of the appropriate dissimilarity (or similarity) measure.

1 The measurement scales of variables

In the theory of measurement four basic scales are distinguished: nominal, ordinal, interval and ratio. Among the four scales of measurement, the nominal is considered the lowest. It is followed by the ordinal scale, the interval scale, and the ratio scale, which is highest. They were introduced by Stevens (1959). The systematic of scales is related to the transformations, which retain the relations of respective scale. This is summarised in Table 1.

One of the basic rules in the measurement theory is the following one: the numbers being the results of the measurement on the stronger (higher) scale can be transformed to the numbers on the weaker (lower) scale. The transformation of values from weaker scale to stronger scale is not permissible, since this means increasing the amount of available information. Anderberg (1973) presents some approximation methods of transformation from weaker scale to stronger scale by using some additional information.

A general and important guideline is that the statistics based on a lower level of measurement can be used for a higher scale of measurement, since permissible functions for higher scale are also permissible for lower scale.

Hand (1996) discusses the problem of relationship between measurement scales and statistics. He presents the major theories of measurement and describes the different kinds of models which may be derived within each theory. He shows in this article several examples, which has been the source of confusion and controversy.

Table 1: The Rules for Scales of Measurement

Scale	Basic Empirical Operations	Allowed Mathematical Transformations	Allowed Arithmetic Operations
Nominal	equal to ($x_A = x_B$), not equal to ($x_A \neq x_B$)	$z = f(x)$, $f(x)$–any one-to-one correspondence function	counting of events (numbers of relations equal to, not equal to)
Ordinal	above and greater than ($x_A > x_B$), smaller than ($x_A < x_B$)	$z = f(x)$, $f(x)$–any strictly increasing function	counting of events (numbers of relations equal to, not equal to, greater than, smaller than)
Interval	above and equality of differences $x_A - x_B = x_C - x_D$	$z = bx + a$ $(b > 0)$, $z \in R$ for all possible values x in R. The zero value on this scale is usually assumed, either arbitrarily or by the convention	above and addition, subtraction
Ratio	above and equality of ratios $(\frac{x_A}{x_B} = \frac{x_C}{x_D})$	$z = bx$ $(b > 0)$, $z \in R_+$ for all possible values x in R_+. The natural origin of the ratio scale is zero (this scale is bounded from the left)	above and multiplication, division

Source: Adapted from: Stevens (1959), p. 25, 27; Walesiak (1995), p. 189-191.

2 Standardisation of variables

Multivariate statistical methods often require that the scales of measurement of all variables are either the same or at least similar (as similar interval and ratio scale are considered as well as nominal and ordinal). In addition, in many multivariate statistical methods, like clustering or multidimensional scaling, one has to standardise the variables.

The purpose of standardisation is to adjust the size (magnitude) and the relative weighting of the input variables (see e.g. Milligan and Cooper (1988), p. 182). The standardisation is used when the variables are measured on interval or ratio scale. In the case of nominal and ordinal scales, standardisation is not necessary, because on nominal and ordinal values such relations as equality of differences and equality of ratios are not permitted.

The only permissible transformations on the interval and ratio scale are linear transformations, thus the standardisation formulas are of the following type (Walesiak (1990)):

$$z_{ij} = bx_{ij} + a \quad (b > 0), \tag{1}$$

where $z_{ij}(x_{ij})$ denotes the value (standardised value) of the j-th variable for the i-th object.
The particular (often used) case of transformation (1) is the one where:

$$b = 1/\sigma, \quad a = -\mu/\sigma, \tag{2}$$

here: μ – location parameter,
σ – spread (scatter) parameter.

This can be also given as:

$$z_{ij} = (x_{ij} - \mu)/\sigma. \tag{3}$$

Therefore, in this case we have general type of standardisation. In this generalisation instead of mean more general location parameter is used and instead of standard deviation more general spread parameter is used.
This type of standardisation leads to standardised variables where for each variable:

- location parameter is equal to 0,
- spread parameter is equal to 1.

Among the possible location and spread parameters are those based on L_p-norm. To derive this, we use the arguments given by Jajuga (1999). First let us start with the case of L_2-norm. It is well know that the location parameter is mean and spread parameter is standard deviation. Let us note that here:

- location parameter (in this case – mean) is the solution to the problem of the minimisation (with respect to μ) $\sqrt{\sum_{i=1}^{n}(x_i - \mu)^2}$,
- spread parameter (in this case – standard deviation) is equal to $\sigma = \sqrt{\sum_{i=1}^{n}(x_i - \mu)^2}/\sqrt{n}$.

So the location parameter is the solution of minimisation problem and spread parameter is the "volume" of the set of observations measured with respect to particular norm (in this case L_2-norm).
It is well known that the location parameters being the solutions of minimisation problem for other cases of L_p-norm are:

- for $p = 1$: median,
- for $p = \infty$: midrange, given as $\mu = 0.5(x_{max} + x_{min})$.

By using the same argument as for L_2-norm, we can propose the general form of spread parameter for L_p-norm $\sigma = \sqrt[p]{\sum_{i=1}^{n} |x_i - \mu|^p} / \sqrt[p]{n}$ (where μ – corresponding location parameter).

Then it is straightforward to present two other particular cases of spread parameter:

- for $p = 1 : \sigma = \sum_{i=1}^{n} |x_i - \mu| / n$ (where μ – median). So this is mean of absolute deviations from median;
- for $p = \infty : \sigma = 0.5(x_{max} - x_{min})$. Therefore, this is half of range.

By assuming different norms, we get different possible standardisation formulas.

The another particular types of standardisation are:

- unitisation, where as location parameter mean is taken and as spread parameter range is taken,
- unitisation with zero minimum, where as location parameter minimal value of variable is taken and as spread parameter range is taken.

Another standardisation types are so called quotient transformations being cases of (2) where:

$$b = 1/x_{0j}, \quad a = 0, \tag{4}$$

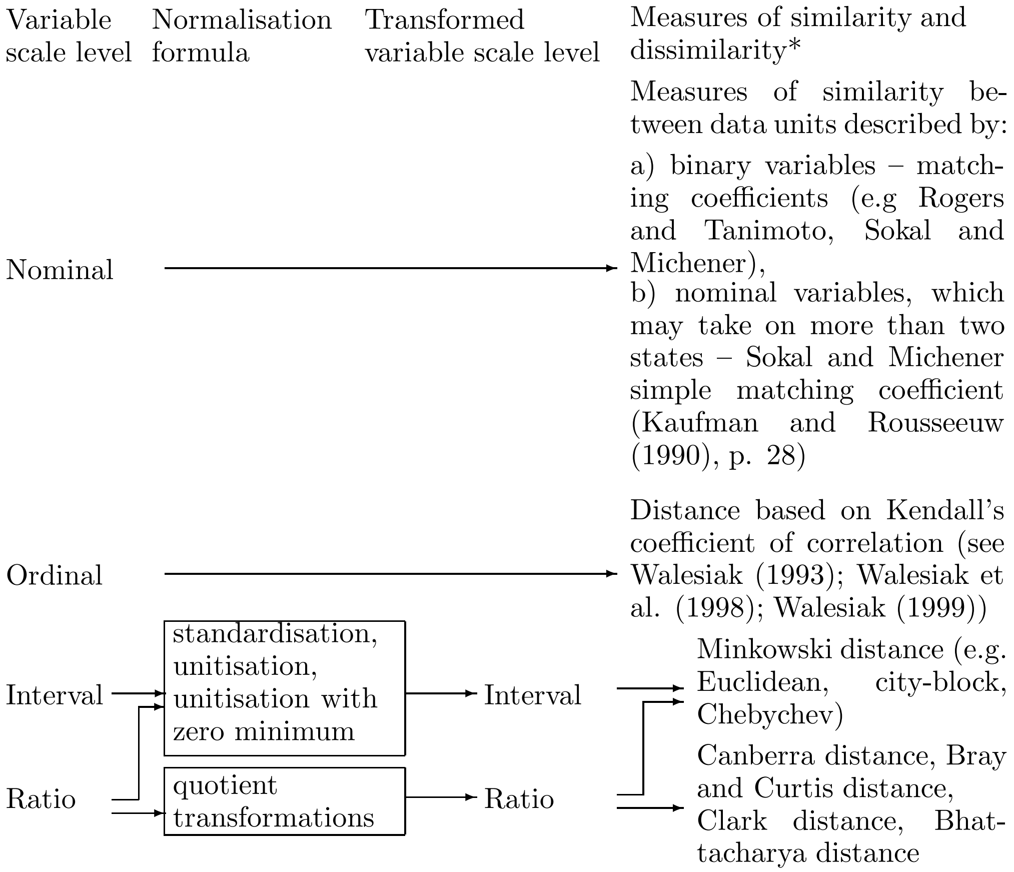
where x_{0j} denotes normalising value, for which the following cases are met in practice: standard deviation, range, maximal value of variable, mean, $x_{0j} = \sum_{i=1}^{n} x_{ij}$ or $x_{0j} = [\sum_{i=1}^{n} x_{ij}^2]^{0.5}$.

The purpose of the standardisation is to equalise the size (magnitude) of variables. This is possible only if the same zero unit for all variables are used. The quotient transformation can be used only if all variables are measured on the ratio scale (for which natural zero unit exists). For the interval scale the general standardisation given by (1) and (2) can be used, provided that for each variable arbitrary zero value is determined by the same procedure.

Multivariate statistical interdependence methods (like clustering methods, multidimensional scaling methods) use different standardisation formulas and similarity or dissimilarity measures. The use of these formulas and measures depends on the particular scale of measurement. This is summarised in Figure 1.

Of course, when choosing appropriate standardisation formulas, one has to take into account not only the measurement scales, but also different characteristics of distribution after standardisation, like e.g. mean, standard deviation, range. The Table 2 shows the characteristics after transformation for several standardisation formulas.

Figure 1: Classification of standardisation formulas and measures of similarity and dissimilarity from the point of view scales of measurement


Variable scale level
Normalisation formula
Transformed variable scale level
Measures of similarity and dissimilarity*
Measures of similarity between data units described by:
a) binary variables – matching coefficients (e.g Rogers and Tanimoto, Sokal and Michener),
b) nominal variables, which may take on more than two states – Sokal and Michener simple matching coefficient (Kaufman and Rousseeuw (1990), p. 28)
Nominal
Distance based on Kendall's coefficient of correlation (see Walesiak (1993); Walesiak et al. (1998); Walesiak (1999))
Ordinal
standardisation, unitisation, unitisation with zero minimum
Interval
Interval
Minkowski distance (e.g. Euclidean, city-block, Chebychev)
Ratio
quotient transformations
Ratio
Canberra distance, Bray and Curtis distance, Clark distance, Bhattacharya distance

* Formulas for measures of similarity and dissimilarity are shown in: Cormack (1971); Cox and Cox (1994), p. 10-11; Wedel and Kamakura (1998), p. 47.

Table 2: Transformed mean, transformed standard deviation and transformed range after standardisation

Formula	Transformed mean	Transformed standard deviation	Transformed range
$(x_{ij}-\overline{x}_j)/s_j$	0	1	r_j/s_j
$(x_{ij}-\overline{x}_j)/r_j$	0	s_j/r_j	1
$[x_{ij}-\min\limits_i\{x_{ij}\}]/r_j$	$[\overline{x}_j-\min\limits_i\{x_{ij}\}]/r_j$	s_j/r_j	1
x_{ij}/s_j	$\overline{x}_j/s_j$	1	r_j/s_j
x_{ij}/r_j	$\overline{x}_j/r_j$	s_j/r_j	1
$x_{ij}/\max\limits_i\{x_{ij}\}$	$\overline{x}_j/\max\limits_i\{x_{ij}\}$	$s_j/\max\limits_i\{x_{ij}\}$	$r_j/\max\limits_i\{x_{ij}\}$
$x_{ij}/\overline{x}_j$	1	$s_j/\overline{x}_j$	$r_j/\overline{x}_j$
$x_{ij}/\sum_{i=1}^n x_{ij}$	$1/n$	$s_j/\sum_{i=1}^n x_{ij}$	$r_j/\sum_{i=1}^n x_{ij}$
$x_{ij}/\sqrt{\sum_{i=1}^n x_{ij}^2}$	$\overline{x}_j/\sqrt{\sum_{i=1}^n x_{ij}^2}$	$s_j/\sqrt{\sum_{i=1}^n x_{ij}^2}$	$r_j/\sqrt{\sum_{i=1}^n x_{ij}^2}$

$\overline{x}_j$, s_j, r_j denotes arithmetic mean, standard deviation and range for j-th variable
Source: Adapted from: Jajuga (1981), p. 33; Milligan and Cooper (1988).

The following remarks should be mentioned:

- unitisation, unitisation with zero minimum and quotient transformation, where the normalising value is range may be useful, since they retain variability (measured via standard deviation) and set up range for all variables equal to 1;
- classical standardisation (z-score) and quotient transformation where normalising value is standard deviation unify variability of all variables, thus here the variability is not the base for the clustering;
- quotient transformations where normalising values are maximal value and "norm" retain the differences in means, standard deviations and ranges;
- quotient transformations where normalising values are mean and sum of observations retain the differences in standard deviations and ranges. It is worth to mention that first formula is used in structural studies where the so-called compositional data are used.

In all discussed standardisation types all variables are treated separately, therefore the standardisation is performed separately for each variable. In such approach the interdependences are not taken into account. Sometimes it is worth to consider the standardisation performed jointly for all considered variables. The only one being the multivariate generalisation of (1) and (2) is given as:

$$\mathbf{z_i} = \mathbf{\Sigma^{-1/2}}(\mathbf{x_i} - \boldsymbol{\mu}), \tag{5}$$

where: $\mathbf{z_i}$ – standardised multivariate observation,
$\mathbf{x_i}$ – multivariate observation,
$\boldsymbol{\mu}$ – mean vector,
$\mathbf{\Sigma}$ – covariance matrix.

This is the case of general "joint" standardisation for L_2-norm. It is worth to see that for L_2-norm we have:

- location vector – mean vector, being solution to the problem of minimisation of the function $\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i} - \boldsymbol{\mu})^{\mathbf{T}}(\mathbf{x_i} - \boldsymbol{\mu})$;
- scatter (spread) matrix – covariance matrix, given as $\mathbf{\Sigma} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{x_i} - \boldsymbol{\mu})(\mathbf{x_i} - \boldsymbol{\mu})^{\mathbf{T}}$.

If one attempts to extend this to other cases of p, this fails because of problems to define inner product (used in the specification of location vector, given above) and outer product (used in the formula for scatter matrix, given above). Both of these are defined only for $p = 2$.

3 Conclusions

The considerations given above confirm the remark provided by Milligan and Cooper that standardisation methods involving division by the range are useful when standardisation is being performed, however one has to take into account all limitations resulting from measurement scales.

The quotient transformations should be used if all variables are measured on ratio scale. Milligan and Cooper (1988), p. 184 propose to add to all values a sufficiently large positive constant if some values of j-th variable are negative. The goal of this operation is to obtain the proportionality property. If some values of the j-th variable are negative, it means that this variable is measured on interval scale. On this scale, the proportionality property is not allowed.

All discussed standardisation formulas, being linear transformations of each variable (separately), retain the skewness and kurtosis of distribution of variables. In addition, for each pair of variables all standardisation formulas retain the value of correlation coefficient.

Use of particular distances depends on the measurement scales of variables after transformations. Some distances, like Canberra, Bray-Curtis, Clark, Bhattacharya may be used only if variables are measured on ratio scale.

Acknowledgements: The research presented in the paper was partly supported by the project KBN 1 H02B 011 16.

References

ANDERBERG, M.R. (1973): Cluster Analysis for Applications. Academic Press, New York, San Francisco, London.

CORMACK, R. M. (1971): A Review of Classification (with Discussion). *Journal of the Royal Statistical Society, Ser. A, (3), 321-367.*

COX, T.F., COX, M.A.A. (1994): Multidimensional Scaling. Chapman and Hall, London.

HAND, D.J. (1996): Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society, Ser. A, (3), 445-492.*

JAJUGA, K. (1981): Metody analizy wielowymiarowej w ilościowych badaniach przestrzennych [Multivariate Methods in Quantitative Spatial Research]. Wrocław University of Economics.

JAJUGA, K. (1999): Some Additions to the Problem of L_p-norm Based Parameters. In: JAJUGA, K., WALESIAK, M. (1999): Klasyfikacja i analiza danych – teoria
i zastosowania. Taksonomia 6. Wrocław University of Economics (in press).

KAUFMAN, L., ROUSSEEUW, P.J. (1990): Finding Groups in Data: an Introduction to Cluster Analysis. Wiley, New York.

MILLIGAN, G.W., COOPER, M.C. (1988): A Study of Standardization of Variables in Cluster Analysis. *Journal of Classification No. 2, 181-204.*

MILLIGAN, G.W. (1995): Issues in Applied Classification: Variable Standardization. *CSNA Newsletter, February, Issue 38.*

STEVENS, S.S. (1959): Measurement, Psychophysics and Utility. In: CHURCHMAN, C.W., RATOOSH, P. (Eds.), Measurement; Definitions and Theories. Wiley, New York.

WALESIAK, M. (1990): Syntetyczne badania porównawcze w świetle teorii pomiaru [Synthetic Comparative Studies in the Light of the Measurement Theory]. *Przegląd Statystyczny z. 1-2, 37-46.*

WALESIAK, M. (1993): Statystyczna analiza wielowymiarowa w badaniach marketingowych [Multivariate Statistical Analysis in Marketing Research]. Wrocław University of Economics, Research Papers No. 654.

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, 185-194.*

WALESIAK, M., DZIECHCIARZ, J., BĄK, A. (1998): Ordinal Variables in the Segmentation of Advertisement Receivers. In: RIZZI, A., VICHI, N., BOCK, H.H. (1998): Advances in Data Science and Classification. Proc. 6th Conf. International Federation of Classification Societies in Rome, Springer, Heidelberg, 655-662.

WALESIAK, M. (1999): Distance Measure for Ordinal Data. *Argumenta Oeconomica* (in press).

WEDEL, M., KAMAKURA, W.A. (1998): Market Segmentation. Conceptual and Methodological Foundations. Kluwer, Boston, Dordrecht, London.