Normalisation with respect to pattern

Iwona Müller-Frączek*

Nicolaus Copernicus University in Toruń, Poland

Abstract

The article presents a new normalisation method of diagnostic variables - normalisation with respect to the pattern. The normalisation preserves some important descriptive characteristics of variables: skewness, kurtosis and the Pearson correlation coefficients. It is particularly useful in dynamical analysis, when we work with the whole population of objects not a sample, for example in regional studies. After proposed transformation variables are comparable not only between themselves but also across time. Then we can use them, for example, to construct composite variables.

keywords: normalisation, standardisation, composite variable, synthetic measure

1 Introduction

In regional studies we often need to compare regions (*objects*) with respect to analyzed *complex* (or *composite*) *phenomenon*. Complex phenomenon is a qualitative phenomenon, that is characterized by some quantitative features, called *diagnostic variables*. Each object is then identified with a point of the multidimensional real space. One of the tools of regional research are *composite variable* (or *synthetic measure*). Composite variable is created to reflect multidimensional points (objets) in the one-dimensional space.

Many advanced methods of constructing synthetic variables have been developed, however the simplest methods are often used in practice. There

*Author's Address: I. Müller-Frączek, Faculty of Economic Sciences and Management, Nicolaus Copernicus University, ul. Gagarina 13 a, 87-100 Toruń, Poland, e-mail: muller@umk.pl

are a lot of such examples (see [2]), one of them is very popular Human Development Index (HDI), which ranks countries into four tiers of socio-economic development. Until 2010 HDI was a uniformly weighted sum of three indicators describing: life expectancy, education, and income per capita.

One of the step of the construction of synthetic measure is bringing diagnostic variables to comparability, called *normalisation* or *standardisation*. Normalisation deprives variables their units and unifies their ranges. There are a lot of normalisation formulas (see [4], [5], [8]). Choosing a proper method is important because normalisation influences on results of object ordering.

The usual stochastic approach can be used to determine parameters needed to normalisation. Then we treat values of variable (observations) as a randomly selected sample of the population. This approach should not be used in regional research, where we work with the whole population of objects. In this case we should use a descriptive (deterministic) approach.

Normalisation formulas are most often given for static analysis, this is for a fixed point in time. A normalisation problem appears when we want to compare situations of regions at several time points. Then the variables should also be comparable across time. To achieve this effect in the stochastic approach one can use all values of variable (both for objects and for time) to determine parameters needed for normalisation. However, this solution is controversial in descriptive approach (see [9]), in addition, it requires incesant conversion of results when later observations occur. In this case we should rather use current observations, so after usual normalisation variables are not comparable across time. Then we can not compare the values of synthetic measures, we can only compare rankings. To solve this problem in the mentioned Human Development Index, the parameters of feature scaling are fixed on levels, that are not related to variable distribution. The levels are justified by substantive reasons. For example, the age of 85 was established as the maximum life expectancy at birth.

The article proposes a new method of feature normalisation - *normalisation with respect to the pattern* (or *pattern normalisation* for short). This name was inspired by the Hellwig's paper (see [3], [1]). The method is consistent with the static approach, but it can be used to compare objects at different time points. The method meets the requirements of normalisation that are suggested in literature (see [4], [6]). It preserves skewness and kurtosis. Moreover, the absolute values of the Pearson correlation coefficients are not changed after normalisation.

In the firs step of the pattern normalisation the nature of variable is determined in the context of analyzed complex phenomenon. We distinguish *stimulants* and *destimulants*. Stimulant is a diagnostic variable that has a

positive impact on the analyzed complex phenomenon, while destimulant negative. In regional research determining the nature of variables is natural. Most often, before normalisation, we turn destimulants into stimulants using their inverse values. Unfortunately, the variables after conversion lose their interpretation and their distributions are changed. In the presented method, we do not converse destimulant before normalisation. Destimulants and stimulants are normalized in different ways.

Determining the nature of variable allows us to choose the most beneficial observation among all values of the variable, maximum for stimulant and minimum for destimulant. We call this value *a pattern*. Next we convert all values with respect to this pattern. After transformation we get comparable variables. All of them are destimulants with clear interpretation. Pattern normalisation can be used in common constructiion of composite variables instead of other methods of normalisation. A possible application is shown in [7].

2 Definition of pattern normalisation

Suppose that a complex phenomenon observed for $n \in N$ regions is analyzed. Assume that we cannot measure this phenomenon, whereas we know a collection of measurable diagnostic variables that characterize it.

Assume that diagnostic variables meet both substantive and statistical requirements, for more details see for example [9]. Let us consider one such variable $x = (x_1, x_2, \ldots, x_n) \in R^n$, which is a stimulant (then we write $x \in S$, S denotes the set of stimulants) or a destimulant ($x \in D$ analogously).

In the first step we choose a pattern - the most beneficial of all values of the variable x. The pattern is unique for all objects and is described by the formula:

$$x^+ = \begin{cases} \max\limits_i x_i & \text{if } x \in S, \\ \min\limits_i x_i & \text{if } x \in D. \end{cases} \tag{1}$$

After specifying the pattern x^+ we can consider a new variable u^+ instead of the variable x given by:

$$u_i^+ = \frac{|x_i - x^+|}{\sum_{j=1}^n |x_j - x^+|} = \begin{cases} \dfrac{x^+ - x_i}{\sum_{j=1}^n (x^+ - x_j)} & \text{if } x \in S, \\ \dfrac{x_i - x^+}{\sum_{j=1}^n (x_j - x^+)} & \text{if } x \in D. \end{cases} \tag{2}$$

The formula (2) determines a certain transformation of initial variable $x = (x_1, x_2, \ldots, x_n)$ into a new variable $u^+ = (u_1^+, u_2^+, \ldots, u_n^+)$. We call it *a normalisation with respect to the pattern.* After this transformation the new variable describes the same aspect of complex phenomenon as described by x. So u^+ is a diagnostic variable of this phenomenon.

The pattern normalisation (2) is not just a technical procedure. New variable has a clear interpretation, u_i^+ specifies the share of distance between the i-th object and the pattern in the total distance of all objects from the pattern. The situation of the i-th object is better when the value of u_i^+ is lower.

The values of variable u^+ characterize the positions of objects in the whole system. This is the same as for other forms of normalisation, but the system is specified in a different way. In the case of the pattern normalisation the system is represented by the sum of distances between objects and pattern, while in common normalisations descriptive characteristics of the distribution of x are used for this purpose.

3 Properties of variable after normalisation

The quantitative description of an immeasurable (qualitative) phenomenon is obtained using synthetic measures. Bringing diagnostic variables to comparability is the first step in the construction of such measure. The pattern normalisation can be used for this purpose.

Assume that diagnostic variables are transformed with respect to their patterns. Then the new set of variables has advantages, which are expected for creating synthetic variables. These properties and some proofs are presented below.

A. Basic properties

A1. All variables after pattern normalisation are unitless, non-negative and limited to interval $[0, 1]$. Because of that, the new set of diagnostic variables contains comparable elements.

A2. Irrespective of the initial nature, variable after the pattern normalisation becomes destimulant. It means that the situation of the i-th object is better when the value u_i^+ is lower. In this sense the pattern normalisation unifies the nature of diagnostic variables.

A3. Transforming of variables does not affect the ordering of objects.

B. Extreme values after pattern normalisation

B1. The variable u^+ can take zero value only for the pattern object. Since the pattern is chosen among values of the variable x, zero value is taken.

$$u_i^+ = 0 \Leftrightarrow x_i = x^+.$$

Proof.

$$u_i^+ = 0 \Leftrightarrow \frac{|x_i - x^+|}{\sum_{j=1}^n |x_j - x^+|} = 0 \Leftrightarrow |x_i - x^+| = 0 \Leftrightarrow x_i = x^+$$

□

B2. The value u_i^+ equals 1 when all objects are patterns except the i-th object. This situation is rather unrealistic.

$$u_i^+ = 1 \Leftrightarrow \bigwedge_{j \neq i} x_j = x^+.$$

Proof.

$$u_i^+ = 1 \Leftrightarrow \frac{|x_i - x^+|}{\sum_{j=1}^n |x_j - x^+|} = 1 \Leftrightarrow |x_i - x^+| = \sum_{j=1}^n |x_j - x^+| \Leftrightarrow \bigwedge_{j \neq i} x_j = x^+$$

□

B3. The maximum value of u^+ depends on the nature of variable x and it is expressed by:

$$\max_i u_i^+ = \begin{cases} \dfrac{\max_i x_i - \min_i x_i}{\sum_{j=1}^n (\max_i x_i - x_j)} & \text{if } x \in S, \\ \dfrac{\max_i x_i - \min_i x_i}{\sum_{j=1}^n (x_j - \min_i x_i)} & \text{if } x \in D. \end{cases}$$

Proof.

If $x \in S$, then:

$$\max_i u_i^+ = \frac{\max_i (x^+ - x_i)}{\sum_{j=1}^n (x^+ - x_j)} = \frac{x^+ - \min_i x_i}{\sum_{j=1}^n (x^+ - x_j)} = \frac{\max_i x_i - \min_i x_i}{\sum_{j=1}^n (\max_i x_i - x_j)}.$$

If $x \in D$, then:

$$\max_i u_i^+ = \frac{\max_i (x_i - x^+)}{\sum_{j=1}^n (x_j - x^+)} = \frac{\max_i x_i - x^+}{\sum_{j=1}^n (x_j - x^+)} = \frac{\max_i x_i - \min_i x_i}{\sum_{j=1}^n (x_j - \min_i x_i)}.$$

□

C. Descriptive characteristics of normalised variables

C1. The mean value of u^+ depends only on the number of objects and is inversely proportional to this number. It is expressed by:

$$\overline{u^+} \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n} u_i^+ = \frac{1}{n}.$$

Proof.

$$\overline{u^+} = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - x^+|}{\sum_{j=1}^{n}|x_j - x^+|} = \frac{1}{n}\frac{\sum_{i=1}^{n}|x_i - x^+|}{\sum_{j=1}^{n}|x_j - x^+|} = \frac{1}{n}$$

□

C2. The variance of u^+ is described by:

$$S^2(u^+) \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n}(u_i^+ - \overline{u^+})^2 = \frac{S^2(x)}{n^2(x^+ - \overline{x})^2}.$$

Proof.

$$S^2(u^+) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|x^+ - x_i|}{\sum_{j=1}^{n}|x^+ - x_j|} - \frac{1}{n}\right)^2$$

If $x \in S$, then:

$$\begin{aligned} S^2(u^+) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{\sum_{j=1}^{n}(x^+ - x_j)} - \frac{1}{n}\right)^2 = \frac{1}{n^3}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{x^+ - \frac{1}{n}\sum_{j=1}^{n}x_j} - 1\right)^2 \\ &= \frac{1}{n^3}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{x^+ - \overline{x}} - 1\right)^2 = \frac{1}{n^3}\sum_{i=1}^{n}\left(\frac{\overline{x} - x_i}{x^+ - \overline{x}}\right)^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\overline{x} - x_i)^2}{n^2(x^+ - \overline{x})^2} \\ &= \frac{S^2(x)}{n^2(x^+ - \overline{x})^2} \end{aligned}$$

The proof is similar when $x \in D$. □

C3. The standard deviation of u^+ depends on the nature of variable x and it is expressed by:

$$S(u^+) \stackrel{def}{=} \sqrt{S^2(u^+)} = \begin{cases} \dfrac{S(x)}{n(x^+ - \overline{x})} & \text{if } x \in S, \\ \dfrac{S(x)}{n(\overline{x} - x^+)} & \text{if } x \in D. \end{cases}$$

C4. The coefficient of variation of u^+ is given by:

$$CV(u^+) \stackrel{def}{=} \frac{S(u^+)}{\overline{u^+}} = \begin{cases} \dfrac{S(x)}{x^+ - \overline{x}} & \text{if } x \in S, \\ \dfrac{S(x)}{\overline{x} - x^+} & \text{if } x \in D. \end{cases}$$

C5. The 3-rd central moment of u^+ is given by:

$$\mu_3(u^+) \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n}(u_i^+ - \overline{u^+})^3 = \frac{\mu_3(x)}{n^3(x^+ - \overline{x})^3}.$$

Proof.

$$\mu_3(u^+) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|x^+ - x_i|}{\sum_{j=1}^{n}|x^+ - x_j|} - \frac{1}{n}\right)^3$$

If $x \in S$, then:

$$\begin{aligned} \mu_3(u^+) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{\sum_{j=1}^{n}(x^+ - x_j)} - \frac{1}{n}\right)^3 = \frac{1}{n^4}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{x^+ - \frac{1}{n}\sum_{j=1}^{n}x_j} - 1\right)^3 \\ &= \frac{1}{n^4}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{x^+ - \overline{x}} - 1\right)^3 = \frac{1}{n^4}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\overline{x} - x^+}\right)^3 \\ &= \frac{\mu_3(x)}{n^3(\overline{x} - x^+)^3} \end{aligned}$$

The proof is similar when $x \in D$. □

C6. The absolute value of the coefficient of skewness does not change after the pattern normalisation:

$$A(u^+) \stackrel{def}{=} \frac{\mu_3(u^+)}{S^3(u^+)} = \begin{cases} -A(x) & \text{if } x \in S, \\ A(x) & \text{if } x \in D. \end{cases}$$

C7. The 4-th central moment of u^+ is given by:

$$\mu_4(u^+) \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n}\left(u_i^+ - \overline{u^+}\right)^4 = \frac{\mu_4(x)}{n^4(x^+ - \overline{x})^4}.$$

Proof.

$$\mu_4(u^+) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{\sum_{j=1}^{n}(x^+ - x_j)} - \frac{1}{n}\right)^4$$

If $x \in S$, then:

$$\begin{aligned}\mu_4(u^+) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{\sum_{j=1}^{n}(x^+ - x_j)} - \frac{1}{n}\right)^4 = \frac{1}{n^5}\sum_{i=1}^{n}\left(\frac{x^+ - x_i}{x^+ - \frac{1}{n}\sum_{j=1}^{n} x_j} - 1\right)^4 \\ &= \frac{1}{n^5}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\overline{x} - x^+}\right)^4 = \frac{1}{n^5}\sum_{i=1}^{n}\left(\frac{\overline{x} - x_i}{x^+ - \overline{x}} - 1\right)^4 \\ &= \frac{\mu_4(x)}{n^3(\overline{x} - x^+)^4}\end{aligned}$$

The proof is similar when $x \in D$. □

C8. The kurtosis of u^+ does not change after the pattern normalisation:

$$K(u^+) \stackrel{def}{=} \frac{\mu_4(u^+)}{S^4(u^+)} = K(x).$$

D. Linear relation between variables after normalisation

Assume that two diagnostics variables x_1, x_2 are transformed with respect to their patterns. Denote by u_1^+ and u_2^+ variables after normalisation.

D1. The covariance between u_1^+ and u_2^+ equals:

$$\begin{aligned}cov(u_1^2, u_2^+) &\stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n}\left(u_{i1}^+ - \overline{u_1^+}\right)\left(u_{i2}^+ - \overline{u_2^+}\right) \\ &= \begin{cases} \dfrac{cov(x_1, x_2)}{n^2(x_1^+ - \overline{x_1})(x_2^+ - \overline{x_2})} & \text{if } x_1, x_2 \in S \text{ or } x_1, x_2 \in D, \\ \dfrac{-cov(x_1, x_2)}{n^2(x_1^+ - \overline{x_1})(x_2^+ - \overline{x_2})} & \text{otherwise.} \end{cases}\end{aligned}$$

Proof.

$$cov(u_1^2, u_2^+) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|x_{i1} - x_1^+|}{\sum_{j=1}^{n}|x_{j1} - x_1^+|} - \frac{1}{n}\right)\left(\frac{|x_{i2} - x_2^+|}{\sum_{j=1}^{n}|x_{j2} - x_2^+|} - \frac{1}{n}\right)$$

Assume that x_1 and x_2 are stimulants. The proof in other cases is similar.

$$\begin{aligned}
cov(u_1^2, u_2^+) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_1^+ - x_{i1}}{\sum_{j=1}^{n}(x_{j1} - x_1^+)} - \frac{1}{n}\right)\left(\frac{x_2^+ - x_{i2}}{\sum_{j=1}^{n}(x_{j2} - x_2^+)} - \frac{1}{n}\right) \\
&= \frac{1}{n^3}\sum_{i=1}^{n}\left(\frac{x_1^+ - x_{i1}}{x_1^+ - \frac{1}{n}\sum_{j=1}^{n} x_{j1}} - 1\right)\left(\frac{x_2^+ - x_{i2}}{x_2^+ - \frac{1}{n}\sum_{j=1}^{n} x_{j2}} - 1\right) \\
&= \frac{1}{n^3}\sum_{i=1}^{n}\left(\frac{x_1^+ - x_{i1}}{x_1^+ - \overline{x_1}} - 1\right)\left(\frac{x_2^+ - x_{i2}}{x_2^+ - \overline{x_2}} - 1\right) \\
&= \frac{1}{n^3}\sum_{i=1}^{n}\left(\frac{\overline{x_1} - x_{i1}}{x_1^+ - \overline{x_1}} \cdot \frac{\overline{x_2} - x_{i2}}{x_2^+ - \overline{x_2}}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \overline{x_1})(x_{i2} - \overline{x_2})}{n^2(x_1^+ - \overline{x_1})(x_2^+ - \overline{x_2})} \\
&= \frac{cov(x_1, x_2)}{n^2(x_1^+ - \overline{x_1})(x_2^+ - \overline{x_2})}
\end{aligned}$$

□

D2. The absolute value of the Pearson correlation coefficient of diagnostic variables is preserved after the normalisation:

$$corr(u_1^+, u_2^+) \overset{def}{=} \frac{cov(u_1^2, u_2^+)}{S(u_1^+)S(u_2^+)} = \begin{cases} corr(x_1, x_2) & \text{if } x_1, x_2 \in S \text{ or } x_1, x_2 \in D, \\ -corr(x_1, x_2) & \text{otherwise.} \end{cases}$$

E. Dynamic approach

Assume that the diagnostic variable x is observed in two periods of time (then we write x^1 and x^2 respectively). For each period we choose a pattern and transform x^1 and x^2 into u^{1+} and u^{2+} according to the formula (2).

E1. The values of variables u^{1+} and u^{2+} are comparable.

Substantiation.

The system is characterized by the sum of distances between objects and the pattern. It changes over time. For given object, if the value of the transformed variable increases over time, this means that the share of distance from this object to the pattern in the sum of all distances increases, so the situation of this object becomes worse (in comparison with the situations of other objects).

4 Summary

The normalisation of diagnostic variables described by formula (2) plays a double role in the construction of synthetic measure. First, it unifies the nature of variables (A2). Secondly, it brings variables to comparability (A1). So, after pattern normalisation diagnostic variables become comparable destimulants. The normalisation with respect to the pattern preserves two important characteristics of the distribution of diagnostic variables - skewness (C6) and kurtosis (C8). Moreover, this conversion does not disrupt linear relation between variables - the absolute value of the Pearson correlation coefficient is not changed (D2). This advantages are expected for normalisations used for bringing variables to comparability.

Unlike other methods the pattern normalisation is not just a technical procedure, it has clear interpretation. However, the major advantage of the pattern normalisation over other normalisation methods appears in dynamic approach. Although the current data are the sole data used to convert variables, the transformed variables are comparable in time (E1).

The normalisation with respect to the pattern seems to be a useful tool in multidimensional comparative analysis. It can be applied whenever variables need to be comparable, for example in the synthetic analysis of complex phenomenon.

The proposed construction can have various modifications, for example we can change the measure of distance or the method of choosing pattern.

References

[1] FANCHETTE, S. (1972) "Synchronic and diachronic approaches in the Unesco project on human resources indicators - Wroclaw taxonomy and bivariate diachronic analysis", UNESCO document, SHS/WS/209, Paris.

[2] FREUDENBERG, M. (2003), "Composite Indicators of Country Performance: A Critical Assessment", OECD Science, Technology and Industry Working Papers, No. 2003/16, OECD Publishing, Paris.

[3] HELLWIG, Z. (1968), "Procedure of Evaluating High-Level Manpower Data And Typology of Countries by Means of the Taxonomic Method", unpublished UNESCO working paper, COM/WS/91, Paris.

[4] JAJUGA, K., WALESIAK, M. (2000), "Stamdardisation of Data Set Under Different Measurement Scales", in *Classification and Information Processing at the Turn of the Millennium. Studies in Classification,*

Data Analysis, and Knowledge Organization, eds. Decker R., Gaul W., Springer-Verlag, Berlin, Heidelberg, 105-112.

[5] MILLIGAN, G.W., COOPER, M.C. (1988), "A Study of Standardization of Variables in Cluster Analysis", *Journal of Classification 5*, 181-204.

[6] MŁODAK, A. (2006), "Multirateral Normalisations of Diagnostic Features", *Statistics In Transition 7(5)*, 1125-1139.

[7] MÜLLER-FRĄCZEK, I. (2017), "Propozycja miary syntetycznej" [Proposition of Synthetic Measure], *Przegląd Statystyczny, 64(4)*, 413-428.

[8] STEINLEY, D. (2004), "Standardizing Variables in K-means Clustering" in *Classification, Clustering, and Data Mining Applications. Studies in Classification, Data Analysis, and Knowledge Organisation*, eds. Banks D., McMorris F.R., Arabie P., Gaul W., Springer, Berlin, Heidelberg.

[9] ZELIAŚ A. (2002), "Some Notes of the Selection of Normalisation of Diagnostic Variables", *Statistics In Transition 5(5)*, 787-802.