Everything you always wanted to know about normalization (but were afraid to ask)

Article *in* Rivista Italiana di Economia, Demografia e Statistica · January 2021

CITATIONS
7

READS
614

2 authors:

Matteo Mazziotta
Italian National Institute of Statistics
46 PUBLICATIONS **916** CITATIONS
SEE PROFILE

Adriano Pareto
Italian National Institute of Statistics
69 PUBLICATIONS **1,211** CITATIONS
SEE PROFILE

EVERYTHING YOU ALWAYS WANTED TO KNOW ABOUT NORMALIZATION (BUT WERE AFRAID TO ASK)[1]

Matteo Mazziotta, Adriano Pareto

1. Introduction

Researchers of social science often use statistical techniques with data that have different units of measurement and ranges (e.g., "Life expectancy" and "Gross national income per capita"). A typical example is the construction of a composite index, where a set of individual indicators of different nature must be aggregated (Salzman, 2003). A common practical problem the researcher faces is how to normalize indicators in order to make them comparable. Unfortunately, most of the reports in the literature describe the main normalization methods (OECD, 2008), but do not explain how to choose the method that best suits the needs of the researcher. So, many researchers choose the normalization method almost solely on the basis of how they want to present the results (e.g., indicators are converted to a common scale with range [0, 1] or to a common scale where a reference is set equal to 100). However, the normalization method has a strong impact on results for two important reasons: (1) it creates a 'correspondence system' between different variables (McGranahan, 1970), (2) it assigns them 'implicit weights' (Booysen, 2002). The 'correspondence system' defines what level of any one variable tends to go with (corresponds to) given levels of other variables (e.g., what level of "Life expectancy" should be found normally with a given level of "Gross national income per capita" or of "Hospital beds per 100,000 inhabitants" and vice versa). These correspondences are particularly important when a non-compensatory approach (i.e., an approach based on the concept of 'unbalance' or disequilibrium among variables) is followed (Casadio Tarabusi and Guarini, 2013)[2]. In such a case, in fact, it is necessary to define what is meant by 'balance' and this definition depends on the normalization method adopted. For example, if indicators are converted to a common

[1] The paper is the result of the common work of the authors: in particular M. Mazziotta has written Sections 2.1-2.3 and 4 and A. Pareto has written Sections 1, 2.4 and 3.

[2] We say that an approach is non-compensatory when it is not full-compensatory. Note that the arithmetic mean is full-compensatory, the geometric mean is partially-compensatory and the minimum is non-compensatory.

scale with range [0, 1], then the set of the maximum values and the set of the minimum values will be considered 'balanced'[3], whereas the set of the mean values could be considered 'unbalanced'. By contrast, if indicators are converted to a common scale where the mean value is set equal to 100, then the set of the mean values will be considered 'balanced'; whereas the set of maximum values and the set of minimum values could be considered 'unbalanced'. Moreover, the range (i.e., the variability) of normalized indicators acts as implicit weight during the aggregation. The wider the minimum and maximum values are apart, the higher the implicit weighting and vice versa. For example, if indicators are converted to a common scale with range [0, 1], then implicit weights are practically the same. By contrast, if indicators are converted to a common scale where the mean value is set equal to 100, but a normalized indicator ranges between 99 and 101 and other ranges between 50 and 200, the composite index will be dominated by the second indicator.

Therefore, an incorrect choice of the normalization method can lead to an unacceptably large degree of distortion of results. This paper discusses the differences among the main normalization methods, and proposes an alternative, denoted as 'Re-scaling with a reference' that combines the advantages of some of them. Some issues on the 'effect' of normalization and suggestions for a correct choice of the normalization method are also reported.

2. Normalization methods and their properties

2.1. Standardization (or transformation in z-scores)

Standardization is the method most commonly used by statisticians. It converts variables to a common scale with a mean of 0 and a standard deviation of 1. For a generic unit *i* and variable *j*, the formula is:

$$y_{ij} = \frac{x_{ij} - \mathrm{M}_{x_j}}{\mathrm{S}_{x_j}} \tag{1}$$

where x_{ij} is the original value of variable *j* for unit *i*, and M_{x_j} and S_{x_j} are, respectively, the mean and standard deviation of variable *j*. If variable *j* has negative polarity[4], then formula (1) can be multiplied by -1. Transformed scores are known as z-scores and most of them (i.e., about 90%) lay between the values −3 and +3,

[3] Note that this is a strong and less plausible assumption, because the minimum and the maximum of a distribution often are 'outliers' (i.e., 'abnormal' values).

[4] The polarity of a variable is the sign of the relation between the variable and the phenomenon to be measured (Mazziotta and Pareto, 2017).

regardless of the shape of the original distribution. They may be further adjusted if calculations yield awkward values. For example, we can multiply each score by 10 and add 100 to obtain positive and more visually manageable scores (Booysen, 2002)[5].

Standardized variables have the same variance and similar (but not equal) range. Standardization has the advantage of 'centering' the variables around a common average and 'normalizing' their variability (Abdi, 2007).

The method allows to compare the values of the units, both in space and time, with respect to the mean and variance of the distribution. So an increase in the standardized value of a given unit, from one period to another, indicates that the original value has increased compared to the new mean and variance (which could also be decreased), but does not necessarily correspond to an increase of the original value. This can be a limitation of the method when different periods have to be compared.

2.2. Re-scaling (or Min-Max method)

Re-scaling is the method most commonly used by sociologists. It converts variables to a common scale with range [0, 1]. For a generic unit *i* and variable *j*, the formula is:

$$y_{ij} = \frac{x_{ij} - \min_i(x_{ij})}{\max_i(x_{ij}) - \min_i(x_{ij})} \tag{2}$$

where x_{ij} is the original value of variable *j* for unit *i*, and $\min_i(x_{ij})$ and $\max_i(x_{ij})$ are, respectively, a minimum and a maximum that represent the possible range of variable *j* (*goalposts*). If variable *j* has negative polarity, then the complement of (2) with respect to 1 is calculated[6]. Also, in this case, transformed scores may be further adjusted to facilitate reading. For example, we can multiply each score by 1,000 in order to obtain values between 0 and 1,000.

Re-scaled variables have the same range and similar (but not equal) variance. Re-scaling has the disadvantage of 'not centering' the variables around a common average. This can be a big problem, if the researcher is interested in constructing a non-compensatory composite index with an unbalance adjustment method (Casadio Tarabusi and Guarini, 2013).

The particularity of the method consists in the possibility of setting the goalposts

[5] In this case we obtain a common scale with a mean of 100 and an approximate range of 70-130.
[6] The 'complement with respect to 1' is the number to add to make 1.

regardless of the values of the variable in a given period[7]. This allows to compare the values of the units, both in space and time, with respect to a common reference (the goalposts) that does not change from one period to another (in contrast to the standardization, where the reference are the mean and variance of each period). So an increase in the normalized value of a given unit, from one period to another, corresponds to an increase of the original value.

2.3. Indicization (or transformation in index numbers)

Indicization[8] is the method most commonly used by economist. It converts variables to a common scale where a reference is set equal to 1. For a generic unit *i* and variable *j*, the formula is:

$$y_{ij} = \frac{x_{ij}}{x_{oj}} \tag{3}$$

where x_{ij} is the original value of variable *j* for unit *i*, and x_{oj} is the reference value (or base) for variable *j* – generally, the maximum, the mean or an external benchmark. If variable *j* has negative polarity, then a non-linear transformation, such as the reciprocal, could be preliminarily applied; however, indicization is recommended only for indicators with positive polarity[9]. Formula (3) can also be adjusted to set the base equal to 100 or 1,000.

Index numbers have the same reference (e.g., the mean), but can have very different range and variance, because they have the same CV[10] of original variables. Therefore, indicization has the disadvantage of 'not normalizing' the variability of variables and introducing implicit weights. This can be a big problem, if the researcher is interested in constructing a composite index with equal or other explicit weights (Mazziotta and Pareto, 2017).

The base of index numbers can be set regardless of the values of the variable in a given period. This allows to compare the values of the units, both in space and time, with respect to a common reference (the base) that does not change in the various periods. So an increase in the index number of a given unit, from one period to another, corresponds to an increase of the original value.

---

[7] Usually, the goalposts are the minimum and maximum of the variable over an extended period of time, in order to take into account its evolution. Alternatively, they can be fixed by experts.

[8] This method is also known as 'Distance from a reference' (OECD, 2008).

[9] A non-linear transformation of variables causes distortions in the data, because correlations among transformed variables are not equal to correlations among original variables.

[10] The coefficient of variation (CV) is a measure of dispersion, often expressed as a percentage, defined as the ratio between standard deviation and mean.

2.4. Re-scaling with a reference (or Constrained Min-Max method)

Re-scaling with a reference is a method that 'normalizes' the variables – similarly to re-scaling – but use a common reference that allows to 'center' them – like indicization. It converts variables to a common scale where a reference is set equal to 0 e the range is 1. For a generic unit *i* and variable *j*, the formula is:

$$y_{ij} = \frac{x_{ij} - x_{oj}}{\max_i(x_{ij}) - \min_i(x_{ij})} \tag{4}$$

where x_{ij} is the original value of variable *j* for unit *i*, $\min_i(x_{ij})$ and $\max_i(x_{ij})$ are, respectively, a minimum and a maximum that represent the possible range of variable *j* (goalposts) and x_{oj} is the reference value for variable *j*. If variable *j* has negative 'polarity', then formula (4) can be multiplied by -1. Transformed scores may be further adjusted as in the previous methods. For example, we can multiply each score by 60 and add 100 to obtain the normalization formula used in the Adjusted Mazziotta-Pareto Index (Mazziotta and Pareto, 2016).

Transformed scores have the same reference (e.g., the mean) and equal range. This allows to have the advantages of index numbers without introducing implicit weights.

Similarly to re-scaling and indicization, the goalposts and the reference can be set regardless of the values of the variable in a given period. This allows to compare the values of the units, both in space and time, with respect to a common reference that remains constant over all periods. So an increase in the normalized value of a given unit, from one period to another, corresponds to an increase of the original value.

3. Comparing normalization methods

3.1. The 'effect' of normalization

To illustrate the effect of normalization on the distributions of individual indicators, we consider the three indicators represented in Figure 1. The first has an exponential distribution (Exp), the second has a normal distribution (Nor) and the third has a Beta distribution (Bet). The indicators have different mean and variance, as they represent the most disparate phenomena. Figure 2 shows the distributions of normalized indicators by 'standardization' (a), 're-scaling' (b), 'indicization' with base=mean set to 100 (c), 're-scaling with a reference' with reference=mean set to 100 and range 60 (d).

Figure 1 – *Individual indicators with different distributions.*

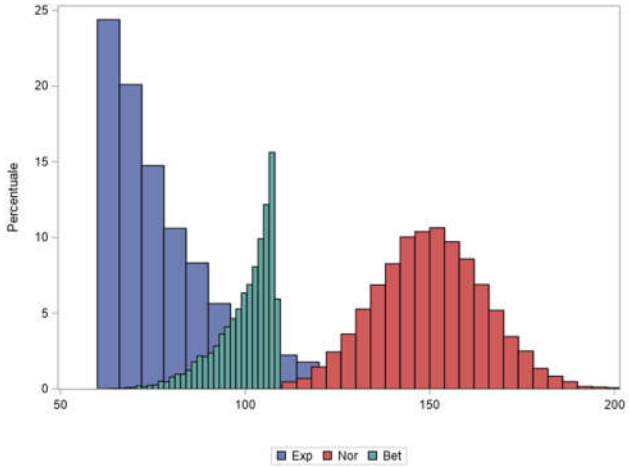
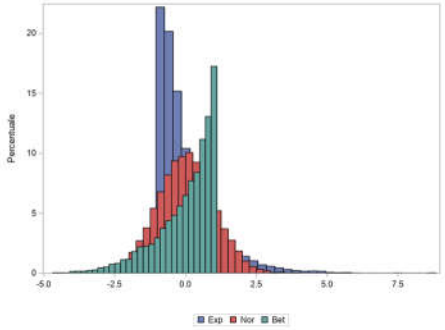
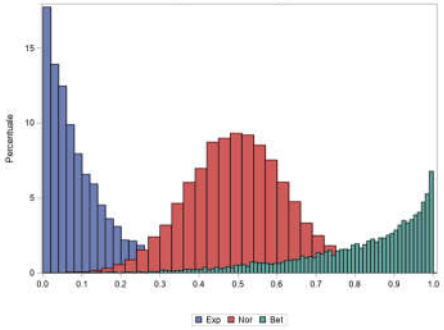


Figure 2 – *Normalized indicators with different distributions.*

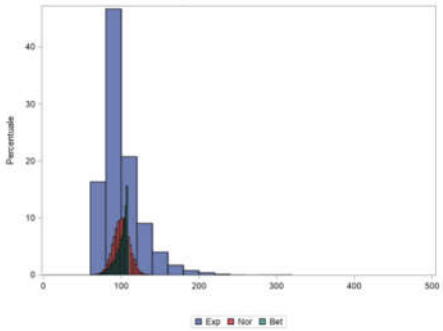
a) standardization



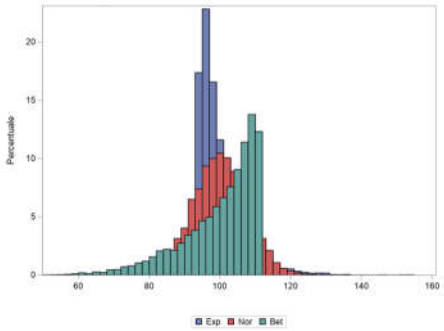
b) re-scaling



a) indicization



b) re-scaling with a reference



As we can see, the distributions of indicators transformed into z-scores (Figure 2.a) are 'centered' around the origin (mean=0) and 'elongated' or 'shortened' to have the same variability (variance=1). So, no implicit weighting is introduced.

Re-scaling also makes the variances more homogeneous (but not equal), bringing all the values into a common interval (Figure 2.b). However, the distributions of indicators are not 'centered' and this leads to the loss of a common reference value, such as the mean. It follows that equal normalized values (i.e. balanced normalized values) can correspond to very unbalanced original values. For example, the normalized value 0.2 for the Exp indicator corresponds to a high original value; whereas for the Nor and Bet indicators it corresponds to a very low original value. Therefore, the use of a simple re-scaling for aggregating individual indicators with an unbalance adjustment method, such as the geometric mean, can lead to biased results. Moreover, the normalized value 0.5 is the mean of the range, but not of distributions, and then it cannot be used as a reference for reading results (e.g., if the normalized value of a given unit is 0.3., we cannot know if its original value is above or below the mean).

Indicization with the mean as a base set to 100 (Figure 2.c) 'centers' all distributions around the mean, but do not 'normalize' their variability (e.g., the range of the Bet indicator is very short, whereas the range of the Exp indicator is very large). So, a significant implicit weighting is introduced when different indicators are compared and aggregated.

Re-scaling with a reference is very similar to standardization when the mean is chosen as a reference (Figure 2.d). In fact, it 'centers' indicators (like indicization) and 'normalizes' them (like re-scaling). Nevertheless, it allows to keep fixed the goalposts and the reference when different periods have to be compared, contrary to standardization. So, the researcher can compare indicators over an extended period of time with respect to a reference, without introducing an implicit weighting.

3.2. The correspondence grid

A correspondence grid emerges when for each indicator the original values that are identified with each level of the common scale are shown on a table. For example, at level 2 of the correspondence grid of standardization are given the original values (for each indicator) that have the standardized value of 2 and that correspond to each other. The result is a list of 'correspondence points' each of which represents a set of original values that will be considered 'balanced'. The correspondence grid must be carefully constructed and evaluated by the researcher, because it can yield an 'artificial' or 'inconsistent' model of balance of original indicators. Table 1 shows the correspondence grids for the four normalizations of Figure 2.

Table 1 – *Correspondence grid for different normalization methods.*

Standardization				Re-scaling			
Scale	**Exp**	**Nor**	**Bet**	**Scale**	**Exp**	**Nor**	**Bet**
2.5	128.5	187.6	119.4	1.0	250.0	209.5	108.4
2.0	118.8	180.1	115.5	0.9	231.0	197.7	103.9
1.5	109.1	172.6	111.6	0.8	212.1	185.8	99.4
1.0	99.4	165.0	107.8	0.7	193.1	174.0	94.9
0.5	89.7	157.5	103.9	0.6	174.2	162.2	90.5
0.0	**80.0**	**150.0**	**100.0**	0.5	155.2	150.3	86.0
-0.5	70.3	142.5	96.1	0.4	136.3	138.5	81.5
-1.0	60.6	135.0	92.2	0.3	117.3	126.7	77.0
-1.5	50.9	127.4	88.4	0.2	98.4	114.8	72.5
-2.0	41.2	119.9	84.5	0.1	79.4	103.0	68.1
-2.5	31.5	112.4	80.6	0.0	60.5	91.2	63.6
Indicization				**Re-scaling with a reference**			
Scale	**Exp**	**Nor**	**Bet**	**Scale**	**Exp**	**Nor**	**Bet**
200	160.0	300.0	200.0	125	158.9	199.3	118.7
180	144.0	270.0	180.0	120	143.2	189.4	114.9
160	128.0	240.0	160.0	115	127.4	179.6	111.2
140	112.0	210.0	140.0	110	111.6	169.7	107.5
120	96.0	180.0	120.0	105	95.8	159.9	103.7
100	**80.0**	**150.0**	**100.0**	**100**	**80.0**	**150.0**	**100.0**
80	64.0	120.0	80.0	95	64.2	140.1	96.3
60	48.0	90.0	60.0	90	48.4	130.3	92.5
40	32.0	60.0	40.0	85	32.6	120.4	88.8
20	16.0	30.0	20.0	80	16.8	110.6	85.1
0	0.0	0.0	0.0	75	1.1	100.7	81.3

There are a number of points of interest in the table. In particular, all normalization methods, except re-scaling, consider 'balanced' the set of mean values. Standardization considers balanced a set of values when they are 'equidistant' from the mean in terms of standard deviations. For example, at level 3 of the correspondence grid are given all the original values that are equal to the mean plus 3 standard deviations. Indicization considers balanced a set of values when they are 'equidistant' from the mean (the base) in percentage terms. For example, at level 200 of the correspondence grid are given all the original values that are double the mean. In this case, also the set of null values is considered balanced, so the transformation in index numbers should be applied only to variables that have an 'absolute zero' point (e.g., "height" and "weight"). Re-scaling with a reference is similar to standardization, but it considers balanced a set of values when they are 'equidistant' from the mean (the reference) with respect to the range. Finally,

classical re-scaling considers balanced the two sets of extreme values and it creates 'artificial' correspondence points (i.e., artificially balanced sets of values) in the middle. For example, the set of the mean values corresponds approximately to the set of normalized values (0.1; 0.5; 0.8), and then it will be considered very unbalanced. The greater the differences between the indicator distributions, the greater the distortion of the correspondence points.

3.3. The implicit weighting

The second issue that the researcher should take into account is the implicit weighting induced by normalization. Suppose we have the following indicators:

- "Life expectancy (years)" (X_1) with a mean of 80 and a standard deviation of 1 (approximate range of 77-83 and CV of 1.25%);
- "Hospital beds per 100,000 inhabitants" (X_2) with a mean of 1,000 and a standard deviation of 200 (approximate range 400-1,600 and CV of 20%).

If we consider a significant change in X_1 from 80 to 82 years (z-score changes from 0 to 2) and a not significant change in X_2 from 1,000 to 1,025 beds (z-score changes from 0 to 0.125), the percentage variations of the two indicators coincide and are equal to 2.5% (both index numbers change from 100 to 102.5). So, normalizing by indicization, the two variations will be considered of equal importance. By contrast, if we consider two variations of equal importance where both indicators increase by 1 standard deviation, that is X_1 changes from 80 to 81 and X_2 changes from 1,000 to 1,200 (both z-scores change from 0 to 1), the percentage variation of X_1 is 1.25% (index number changes from 100 to 101.25), whereas the percentage variation of X_2 is 20% (index number changes from 100 to 120). So, normalizing by indicization, the variation of the indicator with greater CV (X_2) will be considered more important than the variation of the indicator with less CV (X_1).

Table 2 provides an example of aggregating z-scores and index numbers for indicators X_1 and X_2. The aggregation function is a simple arithmetic mean (full compensatory approach). Consider unit 1 and unit 5. In unit 1, X_1 is 1.5 standard deviations below the mean (78.5 years), and X_2 is 1.5 standard deviations above the mean (1,300 beds). Conversely, in units 5, X_1 is 1.5 standard deviations above the mean (81.5 years), and X_2 is 1.5 standard deviations below the mean (700 beds). If we assign the same importance to the indicators, the two units are in a similar situation and therefore must have the same position in a ranking according to the mean of normalized values. However, if we use index numbers, X_2 has a greater weight than X_1 in the computation of the mean, and unit 1 obtains a greater score

than unit 5 (114.1 vs. 85.9) because it has the higher value (1,300) on X_2. On the contrary, if we use z-scores, the two units have the same score (0.0).

Table 2 – *Example of implicit weighting*

Unit	Original indicators		Z-scores			Index numbers		
	X_1	X_2	X_1	X_2	Mean	X_1	X_2	Mean
1	78.5	1,300	-1.5	1.5	0.0	98.1	130.0	114.1
2	80.5	1,100	0.5	0.5	0.5	100.6	110.0	105.3
3	80.0	1,000	0.0	0.0	0.0	100.0	100.0	100.0
4	79.5	900	-0.5	-0.5	-0.5	99.4	90.0	94.7
5	81.5	700	1.5	-1.5	0.0	101.9	70.0	85.9
Mean	**80.0**	**1,000**	**0.0**	**0.0**		**100.0**	**100.0**	
Std	**1.0**	**200**	**1.0**	**1.0**		**1.25**	**20.0**	
CV	**1.25**	**20**				**1.25**	**20.0**	

This simple example shows that indicization makes indicators independent of the unit of measurement, but not of their variability (original values and index numbers have the same CV). The higher the CV, the greater the weight, in terms of normalized values, on the aggregation function. Therefore, in order to assign the same 'importance' to each indicator, the researcher should apply a normalization method that also makes the indicators independent of the variability.

4. Conclusions

Values measured with different units of measurement alone do not explain so much, as each value is meaningful only relative to the mean and variability of the distribution. Therefore, how to compare a life expectancy of 82 years with 800 hospital beds per 100,000 inhabitants? Values from different distributions can be normalized in order to provide a way of comparing them that includes consideration of their respective distributions. This is normally done by transforming the values into z-scores which are expressed as standardized deviations from their means (Abdi, 2007). Nevertheless, standardization is not the best method for comparisons over different periods. In this paper, we propose a similar normalization method, denoted as 'Re-scaling with a reference' (or 'Constrained Min-Max method') that can be used when different periods have to be compared.

It is good to specify that, in general, the perfect normalization method does not exist. Each method has strengths and weaknesses and the choice depends on the aims of the research and/or on the aggregation function used for constructing the

composite index. The paper shows that when choosing the normalization method an implicit weighting must always be avoided and, above all, a realistic 'correspondence grid' (not artificial or meaningless) must be constructed in order to consider a correct 'balancing model' of the values.

For this reason, many composite indices based on the classical Min-Max method should be revised. This is the case of the Human Development Index - HDI (UNDP, 2019), where however the goalposts are 'reasoned' and have been set by experts. In other cases, as in the composite indices summarizing the SDGs, published recently by Istat (Istat, 2020), the computation procedure is based on a re-scaling with 'observed' goalposts and percentage variations of indicators with different CV are considered of equal importance (-80%; +80%). This can lead to strong distortions of the 'balancing model' of indicators and, therefore, to incorrect or misleading results. In this regard, we must remember that the first criterion to be followed in the construction of a composite index (as in any statistical model) is the *principle of parsimony* (Mazziotta e Pareto, 2020). This principle states that the composite index must be as simple as possible, to allow an easy interpretation of results, both in space and time. In order to construct a composite index as simple as possible, the processing to be performed on the data must be reduced to the minimum necessary. Therefore, only one normalization method must be applied to the data matrix and no further transformation of the obtained scores should be carried out, as they are already normalized (for example, it does not make sense to calculate index numbers on values already normalized with re-scaling, because the percentage variations of these values have no meaning).

In conclusion, the construction of a composite index must follow a precise work paradigm and international literature is unanimous in this sense. Methodological shortcuts or even fanciful approaches, such as normalizing data several times, are absolutely to be avoided since a composite index has a great responsibility: measuring multidimensional phenomena to better understand the reality.

References

ABDI H. 2007. Z-scores. In SALKIND N. (Ed) *Encyclopedia of Measurement and Statistics*. Thousand Oaks: Sage.

BOOYSEN F. 2002. An overview and evaluation of composite indices of development, *Social Indicators Research*, Vol. 59, pp. 115-151.

CASADIO TARABUSI E., GUARINI G. 2013. An Unbalance Adjustment Method for Development Indicators, *Social Indicators Research*, Vol. 112, pp. 19-45.

ISTAT 2020. *Rapporto SDGs 2020*. Roma: Istituto nazionale di statistica.

MAZZIOTTA M., PARETO A. 2016. On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena, *Social Indicators Research*, Vol. 127, pp. 983-1003.
MAZZIOTTA M., PARETO A. 2017. Synthesis of Indicators: The Composite Indicators Approach. In MAGGINO F. (Ed) *Complexity in Society: From Indicators Construction to their Synthesis*, Social Indicators Research Series 70, Cham: Springer, pp. 159-191.
MAZZIOTTA M., PARETO A. (a cura di) 2020. *Gli indici sintetici*. Torino: Giappichelli.
MCGRANAHAN D. 1970. The interrelations between social and economic development, *Social Science Information*, Vol. 9, pp. 61-77.
OECD 2008. *Handbook on Constructing Composite Indicators. Methodology and user guide*. Paris: OECD Publications.
SALZMAN J. 2003. *Methodological Choices Encountered in the Construction of Composite Indices of Economic and Social Well-Being*. Ottawa: Center for the Study of Living Standards.
UNDP (2019). *Human Development Report 2019*. New York: United Nations Development Programme.

SUMMARY

Everything you always wanted to know about normalization (but were afraid to ask)

The solution to the problem of normalization of variables with different units of measurement is of primary interest in data analysis. Most of the reports in the literature present a wide variety of normalization methods, but do not explain how to choose the 'right' method. Researchers cannot avoid the question simply by choosing the normalization on the basis of how they want to present the results, as each method has its pros and cons. In this paper, a comparison among the main normalization methods is presented and an alternative method, denoted as 'Re-scaling with a reference' is proposed. Some issues on the 'effect' of normalization and suggestions for a correct choice of the normalization method are also reported.

Matteo MAZZIOTTA, Istat, mazziott@istat.it
Adriano PARETO, Istat, pareto@istat.it