DATA NORMALIZATION FOR AGGREGATING TIME SERIES: THE CONSTRAINED MIN-MAX METHOD[1]

Matteo Mazziotta, Adriano Pareto

1. Introduction

In recent years, composite indices have been increasingly adopted by many institutions for providing a measurement of multidimensional phenomena, such as well-being, development, poverty and so on. Examples of well-known composite indices are the Human Development Index (HDI) created by the United Nations Development Programme (UNDP, 1990; 2010) and the Canadian Index of Well-being (CIW) produced by the University of Waterloo (Michalos *et al.*, 2011).

In both cases, a set of time series of individual indicators with different units of measurement and ranges (e.g., “Life expectancy at birth” and “Gross national income per capita”) are aggregated into a single composite index for various geographical areas. This procedure involves several problems. In particular, it is necessary to normalize the data to make them comparable and the method used must be time-independent, so as not to change the past data every time a new observation is added.

In the HDI individual indicators are converted to a common scale with range [0, 1] by the Min-Max method, whereas in the CIW individual indicators are converted to a common scale where the 1994 value (base value) is set to 100 by ‘indicization’ (i.e., transformation in index numbers).

In this paper, we show that both these normalization methods have some weaknesses and we use an alternative method, the “Constrained Min-Max Method”, that combines the strengths of the two methods, without having their limitations. The method is a generalization of the normalization formula used in the Adjusted Mazziotta-Pareto Index (Mazziotta and Pareto, 2016). An empirical comparison among the three normalization methods is also presented, where the time series of two well-being indicators of the CIW are normalized and then aggregated.

[1] The paper is the result of the common work of the authors: in particular M. Mazziotta has written Sections 1-2 and A. Pareto has written Sections 3-5.

2. The traditional methods

Let x_{ij}^t be the value of individual indicator *j*, for unit *i*, at time *t* (*j*=1, …, *m*; *i*=1, …, *n*; t=1, ..., *k*). We want to build, for each unit, a composite index CI_i^t that summarizes the trend of the individual indicators over time. If individual indicators have different measurement units, the data must be normalized in order to make them comparable both across units and over time.

The first solution that might come to mind is to transform the data in *z*-scores by the classical standardization. However, if a standardization is done over time, adding a new observation would change the mean and variance of the time series and then the data would have to be standardized every time. Furthermore, the data concerning different units would not be easily comparable since the mean of the time series is a reference difficult to interpret. For this reason, the Min-Max method or indicization are usually used.

The Min-Max method is most used by sociologists. It normalizes indicators to have an identical range [0, 1]. For a generic unit *i* and indicator *j* at time *t*, the normalized value is:

$$y_{ij}^t = \frac{x_{ij}^t - \min_{it}(x_{ij}^t)}{\max_{it}(x_{ij}^t) - \min_{it}(x_{ij}^t)}$$

where $\min_{it}(x_{ij}^t)$ and $\max_{it}(x_{ij}^t)$ are, respectively, a minimum and a maximum that represent the possible range of indicator *j* (*goalposts*). They can be calculated across all the units over time or can be fixed by the researcher. However, if new data exceed the selected range, the normalization parameters should be updated in order to avoid values out of the range [0, 1], and normalized values would have to be recalculated for all series.

The Min-Max method normalizes the range of indicators, but it does not ‘centre’ them with respect to a base value and this leads to the loss of a common reference, such as the mean (Mazziotta and Pareto, 2021). Indeed, the normalized value 0.5 is the mean of the range, but not of the 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[2] is the method most used by economist. It measures the relative position of a given value from a reference (Tarantola, 2008). For a generic unit *i* and indicator *j* at time *t*, the normalized value (also called index number) is:

[2] This method is also known as ‘Distance from a reference’ (OECD, 2008).

$$y_{ij}^{t} = \frac{x_{ij}^{t}}{x_{oj}} 100$$

where x_{oj} is the reference (or base) value for indicator *j* (e.g., a mean or the value for a given year).

Indicization 'centre' normalized indicators with respect to a common reference (set equal to 100), but it does not normalize their range, because they have the same CV[3] of original indicators. Therefore, it introduces implicit weights which can affect the aggregation (Mazziotta and Pareto, 2020). The wider the minimum and maximum values are apart, the higher the implicit weighting and vice versa (Booysen, 2002). Therefore, if indicators are normalized by indicization with base 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.

3. The constrained Min-Max method

The Constrained Min-Max method is an alternative method that 'normalizes' indicators – similarly to the Min-Max method – but uses a common reference that allows to 'centre' them – like indicization. It converts indicators to a common scale where a reference is set equal to 0 e the range is 1. For a generic unit *i* and indicator *j* at time *t*, the normalized value is:

$$y_{ij}^{t} = \frac{x_{ij}^{t} - x_{oj}}{\max_{it}(x_{ij}^{t}) - \min_{it}(x_{ij}^{t})} \tag{1}$$

where $\min_{it}(x_{ij}^{t})$ and $\max_{it}(x_{ij}^{t})$ are, respectively, a minimum and a maximum that represent the possible range of indicator *j* (*goalposts*) and x_{oj} is the reference value for indicator *j*.

Normalized indicators have the same reference (e.g., the value for a given year) and equal range. This allows to have the advantages of indicization (normalized values are easier to interpret) without introducing implicit weights. Moreover, transformed scores may be further adjusted if calculations yield awkward values. Finally, if new data exceed the selected range, the comparability across time is

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

maintained and the parameters of formula (1) do not need to be updated.

4. An empirical comparison

As is known, the CIW is a composite index calculated annually, composed of eight domains that measures change in the wellbeing of Canadians over time (Michalos *et al.*, 2011). Let us consider the following two indicators taken from the "Education" domain of the CIW for the 1994 to 2008[4] period:

- "Ratio of childcare spaces to children aged 0 to 5 years of age" (X_1) with a mean of 15.4 and a standard deviation of 2.72 (CV of 17.7%);
- "Average of 5 social and emotional competence scores for 12 to 13 year olds" (X_2) with a mean of 3.2 and a standard deviation of 0.05 (CV of 1.5%).

We chose these two indicators because they have opposite trends over time and very different CVs.

Table 1 – *Comparing normalization methods.*

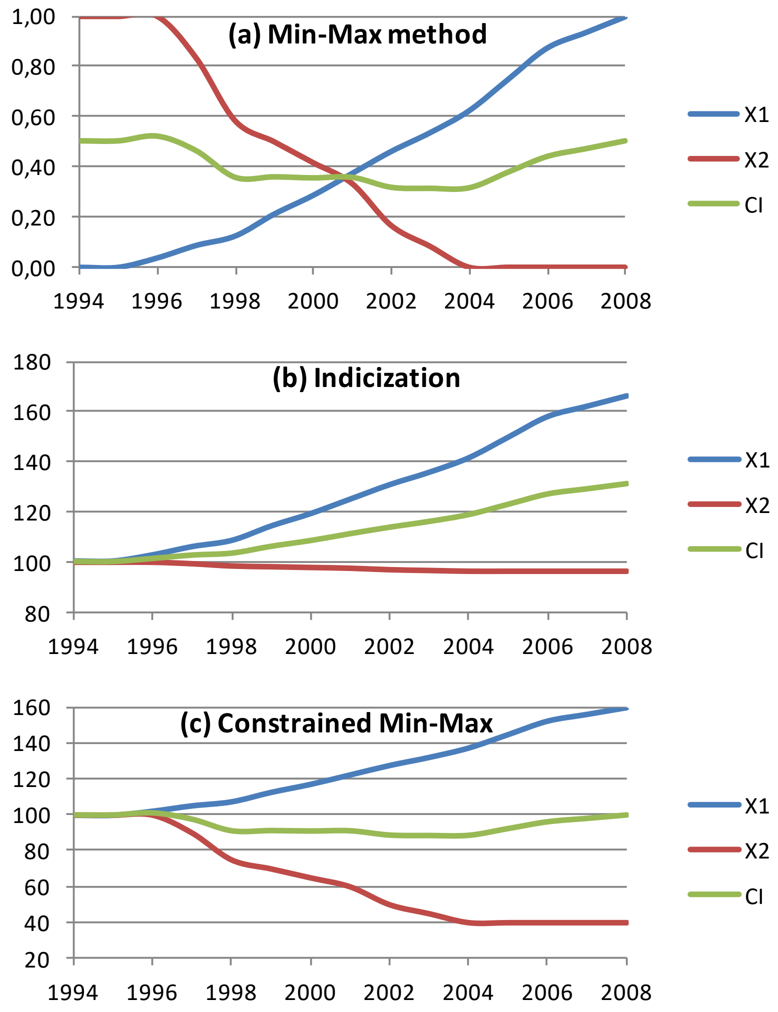
Year	Original data		(a) Min-Max method			(b) Indicization			(c) Constrained Min-Max		
	X_1	X_2	X_1	X_2	CI	X_1	X_2	CI	X_1	X_2	CI
1994	12.0	3.25	0.00	1.00	0.50	100.0	100.0	100.0	100.0	100.0	100.0
1995	12.0	3.25	0.00	1.00	0.50	100.0	100.0	100.0	100.0	100.0	100.0
1996	12.3	3.25	0.04	1.00	0.52	102.5	100.0	101.3	102.3	100.0	101.1
1997	12.7	3.23	0.09	0.83	0.46	105.8	99.4	102.6	105.3	90.0	97.6
1998	13.0	3.20	0.13	0.58	0.35	108.3	98.5	103.4	107.5	75.0	91.3
1999	13.7	3.19	0.21	0.50	0.36	114.2	98.2	106.2	112.8	70.0	91.4
2000	14.3	3.18	0.29	0.42	0.35	119.2	97.8	108.5	117.3	65.0	91.1
2001	15.0	3.17	0.38	0.33	0.35	125.0	97.5	111.3	122.5	60.0	91.3
2002	15.7	3.15	0.46	0.17	0.31	130.8	96.9	113.9	127.8	50.0	88.9
2003	16.3	3.14	0.54	0.08	0.31	135.8	96.6	116.2	132.3	45.0	88.6
2004	17.0	3.13	0.63	0.00	0.31	141.7	96.3	119.0	137.5	40.0	88.8
2005	18.0	3.13	0.75	0.00	0.38	150.0	96.3	123.2	145.0	40.0	92.5
2006	19.0	3.13	0.88	0.00	0.44	158.3	96.3	127.3	152.5	40.0	96.3
2007	19.5	3.13	0.94	0.00	0.47	162.5	96.3	129.4	156.3	40.0	98.1
2008	20.0	3.13	1.00	0.00	0.50	166.7	96.3	131.5	160.0	40.0	100.0

Table 1 shows the original values of the indicators and the normalized values by: (a) Min-Max method, (b) indicization, (c) constrained Min-Max method. In (a) and (c) the goalposts are the minimum and the maximum over time; in (b) and (c)

[4] Since 2009, these indicators have been replaced and then there is no more recent data.

the base/reference is the 1994 value. In order to make it easier to compare (b) and (c), we multiplied formula (1) by 60 and we added 100[5].

Figure 1 – *Comparing normalization methods.*



[5] Note that this is the normalization method used in the Adjusted Mazziotta-Pareto Index (Mazziotta and Pareto, 2016).

Finally, for each normalization method, a composite index (CI) was calculated by a simple arithmetic mean.

As we can see, from 1994 to 1998, X_2 decreases by 1 standard deviation (from 3.25 to 3.20), whereas X_1 increases by 0.35 standard deviation (from 12 to 13), thus the variation of X_2 is wider than that of X_1.

However, if indicators are normalized by indicization, X_1 changes from 100 to 108.3 (+8.3%) and X_2 changes from 100 to 98.5 (-1.5%). So, normalizing by indicization, the variation of the indicator with greater CV (X_1) is considered more important than the variation of the indicator with less CV (X_2).

In contrast, if indicators are normalized by the constrained Min-Max method, X_1 changes from 100 to 107.5 (+7.5%) and X_2 changes from 100 to 75 (-25%). consistently with the different variability of the two indicators. The same happens with the Min-Max method, but in this case the reference (or base value) is lost and reading the results is more difficult.

In Figure 1, the line-plots of X_1, X_2 and CI are reported for the three normalization methods. The effect of the different normalization methods on the trend of the composite index is evident. With indicization, CI is increasing over time, since X_1 has an implicit weight greater than X_2. With the classical and the constrained Min-Max method, CI is more stable over time, since X_1 and X_2 have the same weight and the increase of X_1 is offset by the decrease of X_2.

Nevertheless, if a new year of data becomes available, it may happen that the minimum or maximum across units over time, for one or more indicators, changes. In such case, if the indicators are normalized by the classical Min-Max method using the existing goalposts, some values can fall below 0 or above 1. To prevent this, the goalpost should be updated, and CI should be is recalculated across the past years. If instead indicators are normalized by the constrained Min-Max method, CI maintains comparability between the existing and the new data (similarly to indicization).

5. Conclusions

Normalizing data to make them comparable both across units and over time is not a trivial task. The matter can get complicated if new observations are added every year. The most used methods are the Min-Max method and Indicization. The Min-Max method normalizes the variability of indicators, but do not use a common reference to compare them. Indicization uses a common reference to compare indicators but does not normalize their variability.

The constrained Min-Max method combines the advantages of the two methods, as it normalizes the variability of indicators and uses a common reference to compare them. Furthermore, normalized data maintain comparability even when new data are added.

The method is particularly recommended for the normalization of time series of 'unbounded' indicators[6], when implicit weighting is not desired.

Particular attention must be paid to the choice of normalization parameters (*goalposts* and reference value). They must be kept fixed over time and must not be recalculated every time new data are added. In addition, to facilitate reading, the reference value should be within the range defined by the *goalposts*. When the *goalposts* or the reference value will be considered obsolete, it will be sufficient to define new parameters and normalize all the time series again, similarly to the rebasing of index numbers.

References

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

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 F. MAGGINO (eds.), *Complexity in Society: From Indicators Construction to their Synthesis*, Cham: Springer.

MAZZIOTTA M., PARETO A. (a cura di) 2020. *Gli indici sintetici*. Torino: Giappichelli.

MAZZIOTTA M., PARETO A. 2021. Everything you always wanted to know about normalization (but were afraid to ask). *Italian Review of Economics, Demography and Statistics*, Vol. LXXV, 1, pp. 41–52.

MICHALOS, A.C., SMALE, B., LABONTÉ, R., MUHARJARINE, N., SCOTT, K., MOORE, K., SWYSTUN, L., HOLDEN, B., BERNARDIN, H., DUNNING, B., GRAHAM, P., GUHN, M., GADERMANN, A.M., ZUMBO, B.D., MORGAN, A., BROOKER, A.-S., HYMAN, I. 2011. *The Canadian Index of Wellbeing. Technical Report 1.0,* Canadian Index of Wellbeing and University of Waterloo, Waterloo.

OECD 2008. *Handbook on Constructing Composite Indicators. Methodology and user guide*, Paris: OECD Publications.

[6] Indicators can be divided in 'bounded' and 'unbounded'. An indicator is 'bounded' when it ranges between fixed values (e.g., 'Employment rate'). An indicator is 'unbounded' when there are no predetermined upper or lower limits (e.g., 'GDP per capita') (Mazziotta and Pareto, 2017).

TARANTOLA, S. 2008. *European Innovation Scoreboard: strategies to measure country progress over time*. Luxembourg: Office for Official Publications of the European Communities.

UNDP 1990. *Human Development Report 1990: Concept and Measurement of Human Development*. Oxford: Oxford University Press.

UNDP 2010. *Human Development Report 2010*. New York: Palgrave Macmillan.

SUMMARY

Data Normalization for Aggregating Time Series: the Constrained Min-Max Method

This paper presents a method for normalizing data in time series, when variables have different measurement units and they must be aggregated (e.g., for constructing a composite index). The proposed method, denoted as "Constrained Min-Max Method", normalizes the range of variables, similarly to the Min-Max method, but uses a common reference that allows to 'centre' them, as in the case of index numbers. A comparison with the traditional normalization methods is also shown.

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