A balanced normalization for cross-sectional and longitudinal composite indices in Mexico, 2000–2020

Jesús A. Treviño-Cantú
Universidad Autónoma de Nuevo León,
School of Architecture,
Doctoral program in Urban Affairs,
Nuevo Leon, Mexico
E-mail: jtrevino41@hotmail.com

Implicit weighting in the normalization of aggregate variables generates distorted composite indices (CI). This distortion is associated with the uneven range of each distribution, the fact that the maximum and minimum values differ from one variable to another, and the asymmetry of each variable. Controlling implicit weighting is a pending issue. All normalization procedures reviewed in this research counteract the influence of implicit weighting only partially. The balanced standardization proposed in this study achieves the triple purpose of simultaneously matching the ranges between variables, matching the maximum and minimum between them, and controlling the asymmetry in each distribution. This new procedure neutralizes implicit weighting in the cross-sectional or longitudinal aggregation of variables. The variables of educational backwardness in Mexico illustrate this procedure. The methodological proposal of this research is applicable to any subject where space-time normalization is necessary.

Keywords:
normalization procedures,
implicit weighting,
uneven range,
distinct maximum and
minimum values,
asymmetry

Introduction

Normalization, standardization, and transformation should be clarified to avoid confusion in the elaboration of composite indicators. The transformation modifies the shape of the data distribution; therefore, it is appropriate to remove or smoothen the skewness for data manipulation using parametric statistics. Traditional transformation procedures (e.g., square root or logs of the original values) are used to obtain a Gaussian (normal) distribution of values rather than to standardize or normalize them. There are at least four reasons for the confusion between standardization and normalization: i) the conventional use of terms in statistics. 'Standardization' is specially reserved for z values and 'normalization' for all other rescaling procedures; ii) normalization does not 'normalize' in the sense of transforming data in a normal or Gaussian distribution, but it expresses them in abstract numbers; iii) both normalization and standardization entail rescaling data, but with a different approach, as described in this study. The most known (classical) versions of standardization and normalization do not change the shape of the data distribution; and iv) in data mining, in data preprocessing, there is no difference between standardization and normalization (Han et al. 2023).

In this study, the term 'normalization' is used to refer to any procedure rescaling and expressing the data into comparable, pure, abstract, and dimensionless numbers without necessarily changing the shape of the distribution. This definition includes standardization (z score normalization), which follows the conventional use of statistical terms. Eventually, some recent versions of normalization change the distribution shape (normalized in the Gaussian sense). 'Standardization' instead of 'normalization' may also be used if the data mining terminology is specified.

This study addresses the desirable effects to control the implicit weighting of variables added in a composite index (CI). The general objective of this research is to propose and illustrate a balanced spatiotemporal normalization. It is balanced because the procedure controls the implicit weighting in the asymmetry and the uneven length ranges with unequal maximum and minimum values. It is spatiotemporal because the normalized data are suitable for cross-sectional and longitudinal analysis. Normalization procedures can be transversal or cross-sectional (relative) if they refer to a spatial distribution at a moment in time (e.g., one year) or longitudinal (absolute) if they consider a succession of moments (e.g., several years of a period). This research includes an examination of a selection of the most common transversal and longitudinal versions of alternative normalization procedures (Mazziotta–Pareto 2021, 2022; Sojda–Wolny 2020; Walesiak 2018; Dębkowska–Jarocka 2013).

The review and application of current normalization procedures provide the methodological and empirical framework for the proposed method. The educational backwardness variables in the regions and states of Mexico in the 2000 to 2020 period illustrate the methodological contribution of the new procedure.

CIs are necessary for various disciplinary fields. The suggested balanced normalization procedure is applicable to continuous variables of composite indices of all kinds generated for different geographic scales and time periods.

Background

A composite index (CI) is the synthetic way of numerically expressing a complex idea or message about a specific problem without losing sight of the general framework of the phenomenon considered. This synthetic value only makes sense when it is compared with other values in a place at a given time (transverse or cross-sectional comparison) or with itself over time (longitudinal comparison). These comparisons are possible thanks to the normalization of the variables.

Raw data, for example, may be available in percentages; kilometers; persons per room, bedroom, or square kilometer; minutes; micrograms per cubic meter; and rates per thousand or hundred thousand persons. Normalization expresses the original variables in the same abstract or dimensionless unit of measure so that they are comparable and aggregated in a CI. Other essential functions of normalization are to a) invert the negative polarity of the variables of a CI; b) identify outliers that make it difficult to use parametric statistics; c) reduce the adverse impact of data magnitude; d) ensure that the variables are comparable in time and space; and e) balance the implicit weight in unequal ranges, maximum and minimum values between variables, and asymmetry of each variable. This research focuses on this last issue, which has not yet been considered in the current literature.

Until now, there has been no procedure to control the implicit weighting of variables when they are aggregated into a space-time synthetic index. The three characteristics that exert this implicit weighting are the uneven length range of the values between variables, unequal maximum and minimum values between them, and the asymmetry of each distribution.

Uneven length ranges. The range is the distance between the maximum and minimum values of the variable. The uneven range between variables produces distorted indices because they are biased toward the greater ranges (implicit weighting). Between distributions of uneven range, the greater one dominates because low values in some variables are relatively high in others. Example: In the 'constrained min-max normalization' (z_{MP} in this research) of Mazziotta–Pareto (2022), ranges are the same, but the difference in values at the extremes between variables and the asymmetry in each variable remain unsolved problems.

Unequal highest and lowest values. The variables with higher values at the extremes dominate in the aggregate index, even though they have the same range. Low values in variables with high values (e.g., in a data spreading between 50 to 80) will be high for low-value variables (e.g., in a data spreading between 5 to 35), even though the range is equal between variables (30 in the example). Variables with the same maximum and minimum values have the same range, but the contrary is incorrect. Variables with the same range do not necessarily have the same lowest and highest values. Example: In the min-max normalization (z_{MM} in this study), explained in the OECD (2008), the highest and lowest values are the same between variables, ergo, the range is the same, but they keep the asymmetry of the original values.

Asymmetry. Variables with the same range and maximum and minimum values may have different degrees of asymmetry. The asymmetry creates or magnifies the substitution or compensatory effect in the linear aggregation of the variables. The asymmetry may be to the right or left. The first one, also known as positive skewness, shows the concentration of cases at low values, creating a right-tailed distribution.

The mean appears to the right of (greater than) the median in such a way that the sum or average of the variables with the highest value offsets or 'hides' the variables with the lowest value. In left-skewed variables (negative skew), most cases have high values in the original variable. In this second case, the variables do not significantly differentiate the observations and should be omitted from the investigation (Balcerzak 2016).

The skewness correction usually adjusts for kurtosis as well because normalization smooths and obtains a symmetric distribution, without exaggerated peaks or clustered data at the extremes. This correction can be verified with the standardized kurtosis (zkurtosis), whose value, as in the standardized skewness (zskw), must be in the range ±2.

There are two options to correct skewness (and kurtosis). The first option, the traditional option, includes two steps (Gilthorpe 1995): i) 'normalize' the variables using the known transformation procedures, such as logs, square root, the inverse of the variable, Box–Cox or arc sine of the variable, and ii) the transformed variable is normalized by a usual procedure, such as the min-max method applied to the Human Development Index (HDI). The min-max normalization achieves an equal range and matches values at the extremes of the previously normalized data.

The second option to correct the asymmetry is simultaneously transforming and standardizing the variable, as in the 'median absolute deviation from the median' (z_{MAD}) (Leys et al. 2013) or in the procedures analyzed in this investigation. Only in the new procedure suggested in this research are the lowest and highest values equal between variables; therefore, their ranges are equal, and the standardized skewness is acceptable (±2).

The unwritten rule in CIs is that all variables are positive because an increase in the variable is reflected in an increase in the CI. Otherwise, each normalization or aggregation procedure must provide the option to reverse the polarity of the variable.

Selected normalization procedures

This study is focused on normalization procedures used in both cross-sectional (relative or synchronic) and longitudinal (absolute or diachronic) composite indices. This decision rules out indices of ordinal or hierarchical nature that alter the distance between cases, such as Knox (Ricketts et al. 2006), Borda (Dasgupta–Weale 1992), or percentiles (Acharya–Porwal 2020; Flanagan et al. 2011). These indices are for cross-sectional use only. The only possible space-time comparison of the relative indices refers to the shifts between the ranks, without considering absolute increases or decreases over time in the standardized variable. The research also omits indexing based on the 'distance to a reference' or to the value of a specific year because it does not modify the coefficient of variation (CV) of the original values (Mazziotta–Pareto 2022, 2021).

Cross-sectional normalization

a) z_{t0}–normalization (or standardization). Raw data usually have different measures of central tendency (e.g., mean and median), unequal ranges of values, and skewed distributions. Normalization is usually of interest to express the variables in dimensionless measurements without addressing the rest of the important characteristics in the elaboration of composite indices. This procedure centers the mean at zero and the standard deviation at unity. The z_{t0} normalization is only appropriate for normal raw data where the skewness does not distort the meaning of the parameters. This normalization does not eliminate asymmetry, the minimum and maximum values between variables are different, and the ranges between them are uneven.

$$z_C = \frac{x_i - \bar{x}}{\sigma} \quad (1)$$

where x_i= raw value; $\bar{x}$ and σ are the mean and standard deviation, respectively.

b) Normalized boxes (z_Q). Sibley (1987) suggests a two-step normalization to solve the problem of unequal medians. The first step is to center all box plots at zero to visualize the asymmetric spread of values between the variables. This step subtracts the median from the raw value ('Centered Boxes', z_c). The matching of medians between variables provides the same reference to the procedures seeking to control the dispersion of values. The second step is similar to obtaining z values using the classic Formula z = (x_i-mean)/standard deviation, but it replaces the numerator with the 'Centered box' and divides by the interquartile range ('normalized boxes', z_Q). This process centers the medians at zero and rescales data without matching values at the extremes, eliminating unequal ranges between the variables or modifying the original lopsidedness in each of them.

Centered boxes: $z_c = x_i - Me$, where *Me*= Median. (2)

Normalized boxes (z_Q): $z_Q = \frac{x_i - Me}{Q3 - Q1}$ (3)

where *Q1* and *Q3*= the first and third quartiles, respectively.

c) Min-max normalization (z_{MM}). This procedure initially scales the data from 0 to 1. It is also known as the zero unitarization method (ZUM) (Kukuła 2014) or feature scaling in machine learning algorithms (Wikipedia entry). It has been used in human geography since at least the mid-seventies (Smith 1975). This method matches the same maximum and minimum values for all variables, equalizing the range between them, but the asymmetry in the distribution persists if it exists at all. Assume that there are two variables standardized between 0 and 1 in three different cases. In the first case, one variable is symmetrical, and the other is skewed to the right. The second variable will dominate the synthesis of both variables. In the second case, both variables have similar positive skewness. The addition of both variables is also biased because the mean is far away from the center, which divides the distribution into two equal segments (the median). Finally, in the third case, both variables have the same skewness, but in opposite directions, one positive and the other negative. The variable with negative skewness should be eliminated from the calculations unless its polarity is inverted. It measures an opposite concept to that in the composite index. If it is inverted, skewness in both variables should be controlled to avoid the problems described in the two previous cases. Distributions do not need to be perfectly normal (Gilthorpe 1995). A standardized skewness (the z value of skewness, zSkw) of ±2 is acceptable to add variables with parametric statistics in a composite index:

$$zSkw = \frac{Skewness}{Standard\ error\ of\ Skewness} \tag{4}$$

The min-max procedure, originally in the range between 0 and 1, may be expressed in any range by using its general version (Smith 1975; Han et al. 2023; Il Choi 2019):

$$z_{MM} = \left[\left(\left(\frac{x_i - Min}{Max - Min}\right) * (x_{MaxN} - x_{MinN}\right) + x_{MinN}\right] * 100 \tag{5}$$

where x_{minN}= Minimum value in the new scale of x_i; X_{maxN}= Maximum value in the new scale of x_i; x_{min}= Minimum value observed of x_i; x_{max}= Maximum value observed of x_i.

In this research, the formula for min-max (z_{MM}) expresses the values in the range from 70 to 130 for a better comparison with other versions in this investigation or to use aggregation methods that would not be possible with zero values (e.g., geometric mean or logs).

$$z_{MM} = \left(\frac{x_i - Min}{Max - Min} * 60\right) + 70, \tag{6}$$

where *Min* and *Max* are the lowest and highest values of the variable, respectively.

d) Restricted min-max normalization (z_{MP}). The authors only present the longitudinal version of this method, but it can be easily adapted to cross-sectional analysis (Mazziotta–Pareto 2022; Cutillo et al. 2021). The procedure focuses on the equal range and lower variance of the distribution (compared to z_{MM}) without matching the maximum and minimum values between variables or correcting the asymmetry of the raw data. As in z_{MM}, z_{MP} normalization eliminates the unequal range between variables without guaranteeing symmetry in each of them. The procedure solves the unequal range problem and thus partially solves the unequal variance issue. Unfortunately, variance is also related to skewness since the distribution of values may be more loaded toward one end than the other in variables having the same range. The method is similar to z_{MM}, but the reference is not the minimum value but a convenient point for interpretation (e.g., mean, median, or any convenient benchmark value).

$$z_{MP} = \left(\frac{x_i - Ref}{Max - Min} * 60\right) + 70, \text{ where } Ref\text{= Median or mean.} \tag{7}$$

e) Robust normalization (z_{RN}). This method matches medians and box sizes on an interquartile range from 1 to -1 (Brimicombe 2000). The z_{NR} procedure focuses on *Q1*, *Me*, and *Q3* to standardize and normalize the values, but the maximum and minimum extremes are undefined. This procedure partially controls asymmetry, and the uneven range problem between the variables persists. The formula scales *Q1* and *Q3* to 70 to 130, respectively, for comparison purposes. The median is scaled to one hundred.

$$z_{RN} = \left(\frac{x_i - Me}{Me - Q1} * 30\right) + 100, for\ x_i < Me \tag{8}$$

$$z_{RN} = \left(\frac{x_i - Me}{Q3 - Me} * 30\right) + 100, for\ x_i > Me \tag{9}$$

Terms defined as in the previous formulas.

f) Balanced normalization (z_{BN}). The median equals one hundred with the maximum and minimum equal to 130 and 70, respectively. In this procedure, the lowest and highest values, rather than Q1 and Q3 as in Brimicombe (2000), are between –1 (Minimum) and 1 (Maximum), with the median centered at zero. It corrects the asymmetry and solves uneven range problems; the values at the extremes are the same for all the variables. This procedure is an adaptation of robust normalization (z_{RN}), which, in turn, is an evolution of the 'normalized boxes,' zQ (Figure 1). The formula scales z_{BN} values between 70 and 130 to compare results with z_{MP}, Z_{MM} and z_{RN}.

$$z_{BN} = \left(\frac{x_i - Me}{Me - Min} * 30\right) + 100, for\ x_i < Me \tag{10}$$

$$z_{BN} = \left(\frac{x_i - Me}{Max - Me} * 30\right) + 100, for\ x_i > Me \tag{11}$$

Terms defined as in the previous formulas.

In the case of variables with negative polarity (not considered in the case study), there are at least three options: i) to multiply the term in parentheses by -1; ii) to subtract the normalized variable as follows: 200 - z_{BN}, z_{MP}, z_{MM}, or z_{RN}; iii) to manipulate the terms in the formula. Once normalization for positive polarity is conducted, option ii) is the simplest and fastest way to obtain the normalization of variables with negative polarity.

There are two procedures that segment data such as z_{BN}, but they should not be confused with it. Both methods are omitted in this research. The first method is an adaptation of ZUM for variables that increase in a lower range and decrease after reaching a specific point (Kukuła 2014). This normalization is not suitable for composite indices with positive polarity variables (stimulants or benefit variables). An increase in these variables increases the CI. This is not the case for standardized variables with the inverted- u pattern in the ZUM's adaptation.

The second procedure divides the data distribution into three segments: a reference interval and values above and below this interval (upper and lower intervals, respectively). In the reference interval, data are standardized using the 'median absolute deviation from the median' (z_{MAD}) (Leys et al. 2013; Crews–Peralvo 2007).[1] Values outside the reference interval are standardized using z_{MAD} adapted to the corresponding interval (Łuczak–Just 2021). In this procedure, as in the original z_{MAD} or classical z normalization, the maximum and minimum values do not match between variables. The implicit weighting persists in different maximum and minimum values between variables and therefore in their different ranges.

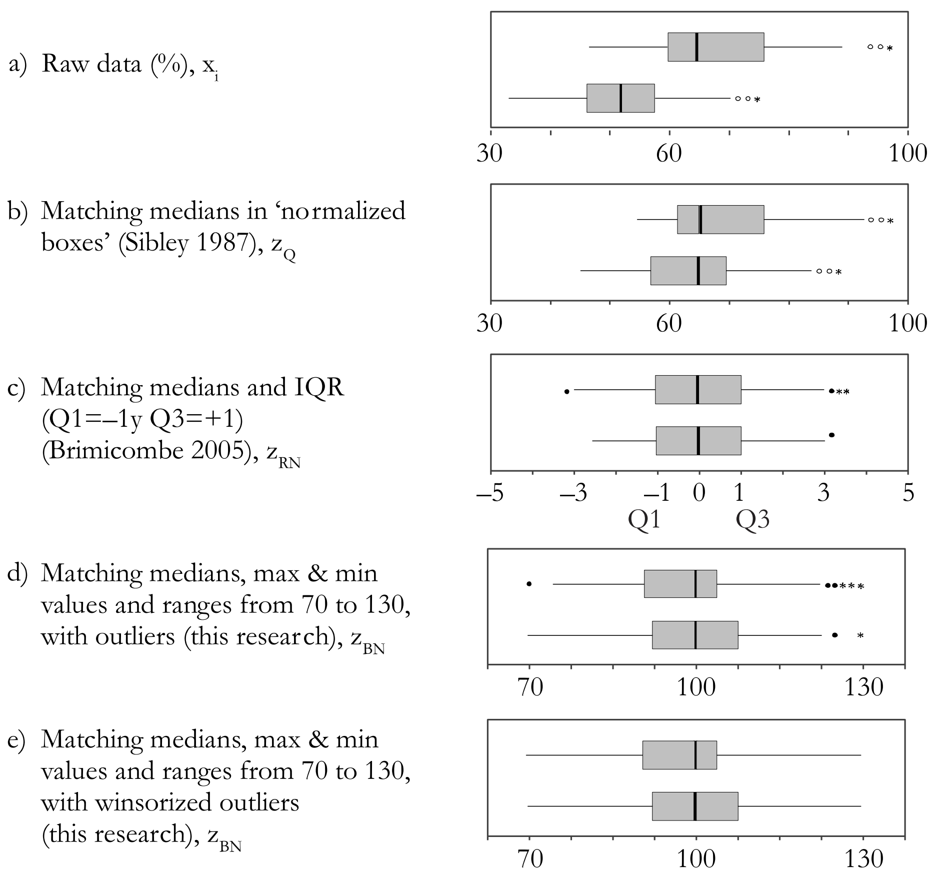
[1] z_{MAD} is also known as modified z score (Iglewicz–Hoaglin 1993), the Weber spatial median (Łuczak–Just 2020) or Oja´s spatial median (Łuczak–Just 2021).

Longitudinal normalization

The data normalization procedure determines whether the composite index is comparable over time (Rinner–Pietropaolo 2021). Raw data refer to the observations of each year, so comparisons over time are not feasible unless the comparison is ordinal (changes in rank or stratum) and discarded in this research. One option to overcome this limitation is piling up the data in an arrangement that stacks the information from one year to the next in a single database, as suggested by Norman (2010, 2015), Exeter et al. (2011) and Norman–Darlington-Pollock (2017). The identification of the maximum and minimum values, median, and range length in this stacked database allows the normalization of each variable considering all the cases in all the years at the same time. Heinrich et al. (2017) refer to this procedure as 'joint standardization.'

Figure 1

Evolution of the balanced normalization procedure (z_{BN})



Note: Elaboration in this study based on Brimicombe (2000).

The longitudinal extension of each cross-sectional index is as follows (z_Q evolves toward robust normalization and balanced normalization):

a) z_{t0} normalization. Mazziotta–Pareto (2022) and OECD (2008) suggest taking the mean and standard deviation of year 0 as a reference, even if the initial cross-sectional or longitudinal distribution is not normal. These parameters remain constant for the entire period. An alternative version is to take as reference the mean and deviation of the stacked data for the entire period (Norman 2010, 2015; Exeter et al. 2011; Norman–Darlington-Pollock 2017). This research does not consider this second version to avoid redundancy and shorten procedures.

b) z_{MM}, z_{MP}, z_{RN} and z_{BN} normalizations. The maximum and minimum values correspond to the stacked data. In all normalizations, except z_{MM}, the reference value can be a specific year, initial (e.g., 2000), or intermediate (e.g., 2010 in the period 2000 to 2020), which remains fixed for all calculations in the period. The example of this research uses the median of all stacked-up years (pooled cross-section data) as a reference. The use of the median instead of the mean is appropriate because it is robust to skewness, and its difference from the mean is not significant in normal distributions. The mean, on the other hand, demands the normality requirement. The normalization z_Q is calculated in an analogous way to these procedures, but the first and third quartiles take the place of the maximum and minimum values.

The balanced normalization (z_{BN}) is on the scale of 100 ± 30 for comparative purposes. A variant of the z_{BN} would be to take as reference (Ref) the average of a specific year, as suggested by Mazziotta–Pareto (2022) for z_{MP}. The Ref could also be the value in each of the cases (e.g., states) of the reference year (e.g., 2010) instead of the average, but that would imply the loss of the hierarchy of the observations in that year (all the 2010 cases would be equal to 100). In this investigation, the Ref for z_{MP} and z_{BN} is the median of the period 2000 to 2020.

Methodology and data

The cross-sectional versions of the selected procedures are first evaluated with box plots and basic descriptive statistics to subsequently address the longitudinal versions applicable to the case study. The statistical significance of changes in the longitudinal analysis is measured with t tests. All calculations and charts are made with *Statgraphics*, v. 16.1.03, *SPSS* v.26.0.0.0 and Microsoft 365 Excel.

The study illustrates the cross-sectional and longitudinal normalization procedures with variables of educational backwardness in the 32 states of Mexico and aggregated in three regions (Figure 2).[2] Educational backwardness is defined by the three variables (components, in CONEVAL terminology) of one out of the five indicators (dimensions) of social backwardness in Mexico: 'illiterate population aged 15 or over'

[2] In addition to the educational backwardness, the social backwardness includes variables on access to health services, housing quality and housing spaces, basic housing services, and household goods (CONEVAL 2007).

(x1); 'population aged 6 to 14 not attending school' (x2); and 'population aged 15 and over with incomplete basic education' (x3). The data come from CONEVAL (2021). The states are grouped into the three large regions of Mexico identified by Angoa et al. (2009).

Figure 2

Regions and states of Mexico

Notes: The states of Mexico: AGS – Aguascalientes; BC – Baja California; BCS – Baja California Sur; CAM – Campeche; CDMX – Mexico City; CHIS – Chiapas; CHIH – Chihuahua; COA – Coahuila; COL – Colima; DGO – Durango; GTO – Guanajuato; GRO – Guerrero; HGO – Hidalgo; JAL – Jalisco; MEX – México State; MICH – Michoacán; MOR – Morelos; NAY – Nayarit; NL – Nuevo León; OAX – Oaxaca; PUE – Puebla; QRO – Querétaro; QR – Quintana Roo; SLP – San Luis Potosí; SIN – Sinaloa; SON – Sonora; TAB – Tabasco; TAM – Tamaulipas; TLA – Tlaxcala; VER – Veracruz; YUC – Yucatán; ZAC – Zacatecas. Adapted from Angoa et al. (2009). Aguascalientes relocated to the Center region. The region for each state is in brackets.

Results

The cross-sectional study included two tasks. The first presents the main characteristics of the normalizations in the regions of Mexico for each year of the period 2000 to 2020 (Table 1). In all the cross-sectional normalizations, in all years of the period, the North reports the least backwardness, followed by the Center and the South, in this order. In classical normalization, the regional hierarchy is the same for all years, and their ranges are uneven. In the min-max, restricted min-max and balanced normalizations, the regions have the same rank, but in the first two, the hierarchy is the same regardless of whether the values increase or decrease from one year to another. Balanced normalization does not allow comparisons because everything is the same from year to year: The values at the extremes (70 and 130), mean (100), stdv (24.5) and range (60) are the same. These results confirm the main limitations of cross-sectional normalization for longitudinal analysis. It is necessary to apply longitudinal normalizations so that the comparisons are adequate, as Mazziotta–Pareto (2022) point out.

Table 1

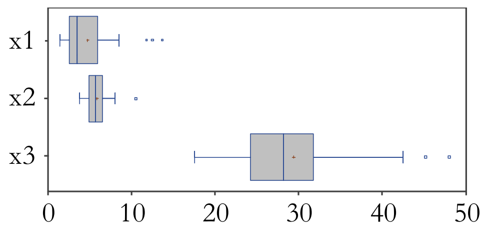
Regions of Mexico. Transversal normalizations, 2000 to 2020 (Raw values x_i in *percentages*)

Raw data (xi), percentages															
Region	x1					x2					x3				
	2000	2005	2010	2015	2020	2000	2005	2010	2015	2020	2000	2005	2010	2015	2020
North	5.0	4.4	3.5	2.7	2.4	7.0	4.4	4.0	3.1	5.3	49.0	42.8	36.8	30.6	25.0
Center	9.6	8.4	6.7	5.2	4.4	7.7	5.1	4.5	3.4	5.8	53.1	46.1	41.2	35.1	29.1
South	15.3	13.8	11.4	9.3	8.3	9.5	5.8	5.5	4.0	6.4	61.0	53.2	47.8	41.8	36.0
Mean	10.0	8.8	7.2	5.8	5.0	8.1	5.1	4.7	3.5	5.8	54.3	47.4	41.9	35.8	30.0
Stdv	4.2	3.9	3.3	2.7	2.4	1.1	0.6	0.6	0.4	0.5	5.0	4.4	4.5	4.6	4.5
Max	15.3	13.8	11.4	9.3	8.3	9.5	5.8	5.5	4.0	6.4	61.0	53.2	47.8	41.8	36.0
Min	5.0	4.4	3.5	2.7	2.4	7.0	4.4	4.0	3.1	5.3	49.0	42.8	36.8	30.6	25.0
Median	9.6	8.4	6.7	5.2	4.4	7.7	5.1	4.5	3.4	5.8	53.1	46.1	41.2	35.1	29.1
z-normalization															
Region	z1					z2					z3				
North	–1.18	–1.16	–1.14	–1.11	–1.07	–1.00	–1.21	–1.04	–1.02	–1.15	–1.08	–1.05	–1.13	–1.13	–1.10
Center	–0.08	–0.11	–0.16	–0.20	–0.26	–0.36	–0.03	–0.30	–0.33	–0.13	–0.25	–0.29	–0.17	–0.16	–0.22
South	1.26	1.28	1.30	1.31	1.33	1.36	1.24	1.35	1.36	1.29	1.33	1.34	1.30	1.30	1.32
Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Stdv	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Range	2.45	2.44	2.43	2.43	2.41	2.37	2.45	2.39	2.38	2.44	2.41	2.40	2.43	2.43	2.42
Min-Max normalization (z_{MM})															
Region	$z1_{MM}$					$z2_{MM}$					$z3_{MM}$				
North	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0
Center	97.1	95.8	94.2	92.7	90.2	86.4	98.7	88.6	87.4	95.1	90.6	89.1	93.8	93.9	91.9
South	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0
Mean	99.0	98.6	98.1	97.6	96.7	95.5	99.6	96.2	95.8	98.4	96.9	96.4	97.9	98.0	97.3
Stdv	24.5	24.6	24.6	24.7	24.9	25.3	24.5	25.1	25.2	24.6	24.9	25.0	24.7	24.7	24.8
Range	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0
Restricted Min-Max normalization (z_{MP})															
Region	$z1_{MP}$					$z2_{MP}$					$z3_{MP}$				
North	41.0	41.4	41.9	42.4	43.3	44.5	40.4	43.8	44.2	41.6	43.1	43.6	42.1	42.0	42.7
Center	68.0	67.2	66.1	65.1	63.5	60.9	69.2	62.4	61.6	66.7	63.7	62.8	65.9	66.0	64.6
South	101.0	101.4	101.9	102.4	103.3	104.5	100.4	103.8	104.2	101.6	103.1	103.6	102.1	102.0	102.7
Mean	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0
Stdv	24.5	24.6	24.6	24.7	24.9	25.3	24.5	25.1	25.2	24.6	24.9	25.0	24.7	24.7	24.8
Range	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0
Balanced normalization (z_{BN})															
Region	$z1_{BN}$					$z2_{BN}$					$z3_{BN}$				
North	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0	70.0
Center	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
South	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0	130.0
Mean	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Stdv	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5
Range	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0	60.0

Note: Calculations by the author. x1: 'illiterate population aged 15 or over'; x2: 'population aged 6 to 14 not attending school'; and x3: 'population aged 15 and over with incomplete basic education'. The standardized box (z_Q) and robust normalization (z_{RN}) are omitted because they evolve to the balanced normalization z_{BN}.

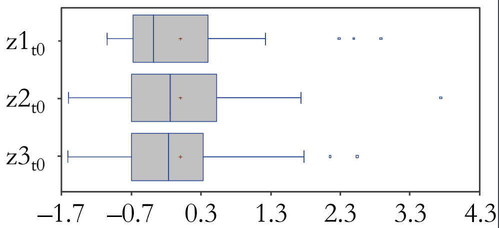
Figure 3

Educational backwardness variables in the Mexican states. Effect of transversal normalization on the raw data (x_i), 2020



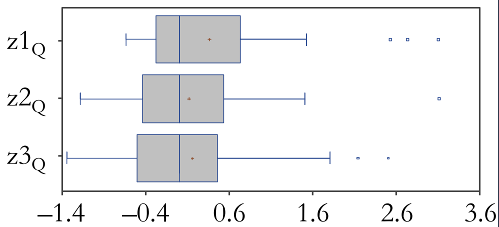
Raw values (x_i). Skewed values with unequal medians, uneven ranges, and different lowest and highest values of the three variables.

	Me	Min	Max	Range	zSke	zKur
x1	3.5	1.40	13.70	12.30	**3.55**	2.27
x2	5.6	3.80	10.50	6.70	**3.65**	5.33
x3	28.2	17.50	48.10	30.60	**2.18**	0.58



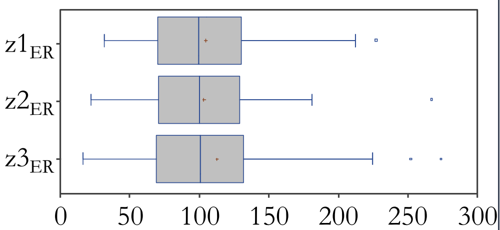
Classic normalization (z_{t0}). Means equal to zero and standard deviation equal to unity. Uneven ranges, outliers in the three variables, same asymmetry of the raw data.

	Mean	stdv	Min	Max	Range	zSke	zKur
$z1_{t0}$	0.00	1.02	–1.05	2.88	3.93	**3.55**	2.27
$z2_{t0}$	0.00	1.02	–1.61	3.74	5.35	**3.74**	5.57
$z3_{t0}$	0.00	1.02	–1.62	2.54	4.16	**2.18**	0.57



Normalized box (z_Q). Equal medians (zero), same asymmetry of the raw data, uneven ranges and outliers in the three variables.

	Me	Min	Max	Range	zSke	zKur
$z1_Q$	–0.43	–0.64	3.10	3.74	**3.55**	2.27
$z2_Q$	–0.40	–1.18	3.11	4.29	**3.74**	5.56
$z3_Q$	–0.66	–1.34	2.50	3.84	**2.18**	0.57

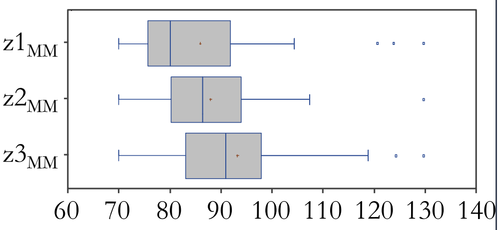


Robust normalization (z_{RN}). Equal medians (100), equal interquartile ranges (70 to 130), uneven ranges. Less asymmetry than previous normalizations, but the skewness problem in the variables x2 and x3 remains. Outliers in the three variables.

	Me	Min	Max	Range	zSke	zKur
$z1_{NR}$	100	30.51	228.48	197.97	1.74	0.46
$z2_{NR}$	100	20.93	268.73	247.80	**3.00**	4.28
$z3_{NR}$	100	15.05	275.51	260.46	**2.41**	0.71

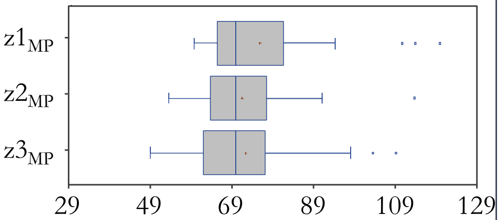
(Figures continue on the next page.)

(Continued.)



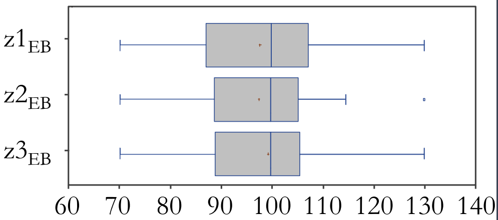
Min-max normalization (z_{MM}). Different medians in the three variables; same extreme values (70 and 130); same range (60). Same asymmetry of the raw data. Outliers in the three variables.

	Me	Min	Max	Range	zSke	zKur
$z1_{MM}$	80	70	130	60	**3.55**	2.27
$z2_{MM}$	86	70	130	60	**3.74**	5.55
$z3_{MM}$	91	70	130	60	**2.18**	0.57



Restricted min-max normalization (z_{MP}). Same median in the three variables; unequal min and max values; same range (60). Same asymmetry of the raw data. Outliers in the three variables.

	Me	Min	Max	Range	zSke	zKur
$z1_{MP}$	70	60	120	60	**3.55**	2.27
$z2_{MP}$	70	54	114	60	**3.74**	5.54
$z3_{MP}$	70	49	109	60	**2.18**	0.57



Balanced normalization (z_{BN}). Same median in the three variables; same min–max values (70 to 130) and same range (60). Outliers only at z2, but with acceptable skewness (±2).

	Me	Min	Max	Range	zSke	zKur
$z1_{EB}$	100	70	130	60	**0.11**	–0.47
$z2_{EB}$	100	70	130	60	**0.15**	1.03
$z3_{EB}$	100	70	130	60	**0.49**	–0.05

Note: Elaboration by the author based on the same data as for Table 1. Variables are the same as in Table 1. Me = median; Min and Max = minimum and maximum values, respectively; Range = max–min; zSkew = skewness/skewness error; zKurtosis = kurtosis/kurtosis error. Subscripts on the z values indicate the following: MM: min-max; MP: Mazziotta-Pareto or restricted MM normalization; BN: balanced normalization. The number in z indicates the standardized educational backwardness variable. Boxplots and standardized skewness and kurtosis values are obtained with Statgraphics, v. 16.1.03.

At the state level, the normalizations provide greater geographical detail of the procedures analyzed and generate the necessary information for the second task of cross-sectional analysis. The 2020 cross-sectional study confirms that the hierarchy is the same for all states, regardless of the normalization used. This research does not provide cross-sectional information in the states for the rest of the period to avoid redundancy since the results are similar to those obtained at the regional level.[3]

The second cross-sectional task, with information from the 32 states of Mexico for 2020, creates box plots that are combined with basic statistics to evaluate the effect of each normalization on the range, the lowest and highest values, and the asymmetry

[3] The information on the transversal and longitudinal normalization is available at the following link: https://github.com/jtrevino41/Pobreza-en-Mx/raw/main/EdBack_Annex.docx

in the three variables of educational backwardness. The normalization procedures, except for z_{RN} and z_{BN}, do not modify the skewness of the original data. In the normalizations with a range from 70 to 130, only the constrained min-max (z_{MP}) does not have exact extreme values. These results show that equal ranges do not guarantee better normalization. It is also necessary to consider the tie of the maximum and minimum value between the variables and the asymmetry in each one of them. Only the balanced normalization (z_{BN}) achieves this triple purpose and confirms that the skewness correction also adjusts the kurtosis (correcting one corrects the other) (Figure 3).

The longitudinal analysis illustrates the normalizations at the level of regions and states. In the regions, in all the normalizations and all the variables of educational backwardness, the maximum corresponds to the South region in the year 2000. The minimum corresponds to the north in 2020, except for the variable z2, whose minimum corresponds to 2015 (Table 2). These results show a decreasing trend of the variables z1 and z3 in all the normalizations. The exception is z2, whose minimum is reached by the North in 2015. The regional arithmetic mean clearly shows these trends in all normalizations. In z1 and z3, the regional average decreases continuously for all the years of the period, while in z2, the trend is reversed from 2015 to 2020. The average of the raw data decreases throughout the period without showing the change in the direction of z2, as detected by longitudinal normalizations. For comparison purposes, the study includes a version of the balanced normalization that uses the national average in 2010 as a reference (z_{BR}), as suggested by the constrained min-max normalization (z_{MP}). The results do not add additional relevant information to that obtained by the balanced normalization with the median of the period as a reference. The possible advantage of z_{BR} is that the reference remains fixed and avoids recalculating the indices when new information is added. This advantage is not convincing because it is necessary to identify the new minimum and maximum value anyway. If these values change, the recalculation of previous values is inevitable, although not very laborious with any computer equipment. The supposed advantage of z_{BR} disappears because this last argument is also valid for the rest of the procedures.

The balanced normalization modifies the original values in a way that is consistent with the rest of the normalizations. Furthermore, as the cross-sectional analysis shows, balanced normalization neutralizes the implicit weight associated with uneven ranges and different maximum and minimum values at the extremes between the variables and the asymmetry of each variable. These characteristics reveal that z_{BN} is an attractive procedure to normalize variables in a transversal or longitudinal way, especially those that are aggregated in a spatiotemporal composite index. This argument is sufficient to describe the longitudinal pattern of the variables at the state level based on z_{BN}.

Table 2

Regions of Mexico. Longitudinal normalizations, 2000 to 2020 (Raw values x_i in *percentages*)

Year	Raw data (x_i): average of the states within each region						z–normalization refered to t_0 (z_{t0}); mean=0 & Stdv= 1 en t_0)					
		North	Center	South	Mean	Stdev		North	Center	South	Mean	Stdev
2000	**x1**	5.03	9.65	**15.26**	9.98	4.18	**$z1_{t0}$**	–1.18	–0.08	1.26	**0.00**	**1.00**
2005		4.36	8.40	13.76	8.84	3.85		–1.34	–0.38	0.90	–0.27	0.92
2010		3.50	6.70	11.44	7.21	3.26		–1.55	–0.78	0.35	–0.66	0.78
2015		2.74	5.24	9.34	5.77	2.72		–1.73	–1.13	–0.15	–1.01	0.65
2020		**2.42**	4.40	8.30	5.04	2.45		–1.81	–1.33	–0.40	–1.18	0.58
2000	**x2**	7.04	7.72	**9.55**	8.10	1.06	**$z2_{t0}$**	–1.00	–0.36	1.36	**0.00**	**1.00**
2005		4.43	5.10	5.82	5.12	0.57		–3.48	–2.85	–2.16	–2.83	0.54
2010		4.05	4.48	5.45	4.66	0.59		–3.84	–3.43	–2.51	–3.26	0.56
2015		**3.08**	3.35	4.03	3.48	0.40		–4.76	–4.50	–3.86	–4.37	0.38
2020		5.32	5.79	6.44	5.85	0.46		–2.64	–2.19	–1.57	–2.13	0.44
2000	**x3**	48.95	53.08	**60.96**	54.33	4.98	**$z3_{t0}$**	–1.08	–0.25	1.33	**0.00**	**1.00**
2005		42.78	46.11	53.22	47.37	4.35		–2.32	–1.65	–0.22	–1.40	0.87
2010		36.78	41.17	47.84	41.93	4.55		–3.52	–2.64	–1.30	–2.49	0.91
2015		30.65	35.08	41.75	35.83	4.57		–4.75	–3.86	–2.52	–3.71	0.92
2020		**25.04**	29.05	36.02	30.04	4.54		–5.88	–5.07	–3.68	–4.88	0.91
Year	Min-Max normalization (zMM) from 70 to 130, Ref=Me for 2000 to 2020						Restricted Min-Max normalization (zMP) from 70 to 130; Ref= Me for 2000 to 2020					
		North	Center	South	Mean	Stdev		North	Center	South	Mean	Stdev
2000	**$z1_{MM}$**	82.2	103.8	**130.0**	105.33	19.54	**$z1_{MP}$**	62.22	83.78	**110.00**	85.33	19.54
2005		79.1	97.9	123.0	100.00	17.99		59.08	77.94	102.99	80.00	17.99
2010		75.1	90.0	112.2	92.41	15.24		55.06	70.00	92.17	72.41	15.24
2015		71.5	83.2	102.3	85.68	12.71		51.52	63.19	82.34	65.68	12.71
2020		**70.0**	79.3	97.5	82.25	11.42		**50.00**	59.27	77.50	62.25	11.42
2000	**$z2_{MM}$**	106.8	113.1	**130.0**	116.63	9.80	**$z2_{MP}$**	85.98	92.31	**109.20**	95.83	9.80
2005		82.6	88.7	95.5	88.92	5.27		61.76	67.94	74.67	68.12	5.27
2010		79.0	83.0	92.0	84.70	5.45		58.21	62.25	71.25	63.90	5.45
2015		**70.0**	72.6	78.8	73.79	3.70		**49.20**	51.76	58.01	52.99	3.70
2020		90.8	95.1	101.2	95.72	4.27		70.00	74.35	80.42	74.92	4.27
2000	**$z3_{MM}$**	109.9	116.8	**130.0**	118.92	8.32	**$z3_{MP}$**	82.02	88.91	**102.08**	91.01	8.32
2005		99.6	105.2	117.1	107.30	7.27		71.72	77.28	89.15	79.38	7.27
2010		89.6	97.0	108.1	98.22	7.59		61.70	69.03	80.17	70.30	7.59
2015		79.4	86.8	97.9	88.02	7.62		51.45	58.85	70.00	60.10	7.62
2020		**70.0**	76.7	88.3	78.35	7.58		**42.08**	48.79	60.42	50.43	7.58

(Table continues on the next page.)

(Continued.)

Year	Balanced normalization (zBN) from 70 to 130; Ref= Me for 2000 to 2020						Restricted balanced normalization (zRB) from 70 to 130; Ref= Country 2010					
		North	Center	South	Mean	Stdev		North	Center	South	Mean	Stdev
2000	**$z1_{BN}$**	88.34	110.33	**130.00**	109.56	17.02	**$z1_{RB}$**	87.60	109.91	**130.00**	109.17	17.32
2005		83.62	105.95	124.74	104.77	16.81		83.07	105.44	124.63	104.38	16.98
2010		77.59	100.00	116.62	98.07	15.99		77.28	98.80	116.34	97.47	15.97
2015		72.27	89.78	109.25	90.43	15.10		72.18	88.99	108.81	89.99	14.97
2020		**70.00**	83.90	105.62	86.51	14.66		**70.00**	83.34	105.10	86.15	14.47
2000	**$z2_{BN}$**	112.23	117.07	**130.00**	119.77	7.50	**$z2_{RB}$**	114.26	118.55	**130.00**	120.93	6.65
2005		88.11	97.03	103.57	96.24	6.34		93.91	102.03	106.59	100.84	5.24
2010		82.99	88.82	100.95	90.92	7.48		87.15	94.84	104.27	95.42	7.00
2015		**70.00**	73.69	82.71	75.47	5.34		**70.00**	74.88	86.78	77.22	7.05
2020		100.00	103.33	107.97	103.77	3.27		103.42	106.37	110.48	106.76	2.90
2000	**$z3_{BN}$**	111.24	117.68	**130.00**	119.64	7.78	**$z3_{RB}$**	111.85	118.08	**130.00**	119.98	7.53
2005		101.61	106.81	117.91	108.77	6.80		102.52	107.55	118.30	109.46	6.58
2010		91.08	98.96	109.51	99.85	7.55		91.92	100.09	110.17	100.73	7.46
2015		80.07	88.02	100.00	89.36	8.19		80.47	88.74	100.97	90.06	8.42
2020		**70.00**	77.21	89.71	78.97	8.14		**70.00**	77.50	90.49	79.33	8.47

Note: Elaboration by the author based on the methodology. Variables are the same as in Table 1. Bold numbers indicate the maximum and minimum value in each variable per period. Calculations based on the original data available at CONEVAL (2021).

The state-level disaggregation reveals movements or changes of direction not detected at the regional or aggregate level. In the states, the absence of negative values in the differences of two out of the three variables with balanced normalization, the ‘illiterate population aged 15 or over’ ($z1_{BN}$) and the ‘population aged 15 and over with incomplete basic education’ ($z3_{BS}$), shows a decreasing evolution in each of the five-year periods and in the period 2000 to 2020. The variable ‘Population from 6 to 14 years old not attending school’ ($z2_{BN}$), on the other hand, reports an increase in the five-year period 2005 to 2010 for the states of COL, CDMX, MOR, QR and SON, not detected in the regional analysis. The trend returns to the downward path in all states in the five-year period 2010 to 2015. The downward direction of $z2_{EB}$ stops in the last five-year period to give rise to a general increase in this variable in all states. The increase in $z2_{BN}$ from 2015 to 2020 is likely due to the sanitary confinement caused by the Covid-19 pandemic. This same explanation applies to the negative values of $z2_{BN}$ in the period 2000 to 2020 for CDMX and NL, where the health policy operated immediately.

All positive or negative changes, by quinquennium or for the entire period, are statistically significant. In all five-year periods, the previous year has a higher average than the following year, except in z2 in the period from 2015 to 2020 (Table 3). The t tests for related data show that the decline in two (z1 and z3) out of the three variables of educational backwardness is statistically significant throughout the period and in each of its five-year periods (Table 3). The changes in z2 are also significant, with the particularity that the statistical significance includes a change in trend in the last five years of the period. However, the result for the period 2000 to 2020 shows that the general trend of z2 is decreasing.

These results show that balanced normalization expresses the variables in the same abstract measurement, controls the characteristics that generate an implicit weight when adding them to a composite index, and allows the statistical analysis of the space-time trend.

Table 3

Educational backwardness variables. T tests for paired samples with balanced standardized data for longitudinal analysis, 2000 to 2020 (n= 32)

Years	Illiterate ($z1_{EB}$)				Not Attending ($z2_{EB}$)				Incomplete ($z3_{EB}$)			
	Mean	Stdv	t	Sig.	Mean	Stdv	t	Sig.	Mean	Stdv	t	Sig.
2000	104.89	12.08	14.99	0.00	107.92	7.38	21.88	0.00	112.73	9.06	34.56	0.00
2005	101.98	12.46			96.91	7.73			105.88	8.76		
2005	101.98	12.46	16.40	0.00	96.91	7.73	4.78	0.00	105.88	8.76	25.64	0.00
2010	97.42	12.68			93.70	8.19			100.04	9.08		
2010	97.42	12.68	15.64	0.00	93.70	8.19	18.74	0.00	100.04	9.08	29.28	0.00
2015	92.18	13.25			83.85	8.86			92.74	9.31		
2015	92.18	13.25	9.82	0.00	**83.85**	8.86	-17.78	0.00	92.74	9.31	25.74	0.00
2020	89.22	13.16			**101.65**	4.85			85.45	9.41		
2000	104.89	12.08	24.11	0.00	107.92	7.38	6.14	0.00	112.73	9.06	39.89	0.00
2020	89.22	13.16			101.65	4.85			85.45	9.41		

Note: Calculations by the author with SPSS v.26.0.0.0. Bold numbers identify the trend reversal in z2. The negative t value registers this change in direction in the trend.

Discussion

All the normalizations in this research are adaptations of classical normalization (z_{t0}). In the normalized boxes (z_Q), the median replaces the mean, and the interquartile range replaces the standard deviation. In min-max (z_{MM}) normalization, the minimum value replaces the mean, and the max-min range replaces the standard deviation. The constrained min-max normalization (z_{MP}) is a version of the min-max normalization where a benchmark (mean, median, or outstanding observation) takes the place of the mean and the max-min range replaces the standard deviation. The robust normalization (z_{RN}) takes into account data above or below the median (Me). In both cases, the numerator is the same: the median takes the place of the mean, $(x_i - Me)$. In data below the median, $(Me - Q1)$ takes the place of the standard deviation; in data above the median, $(Q3 - Me)$ replaces the standard deviation. The balanced normalization (z_{BN}) is similar to z_{RN}, except that the minimum and maximum values take the place of the first and third quartiles, respectively. In $z_{BN,}$ the numerator is the same for values above or below the median: $(x_i - Me)$; in the denominator, $(Me - Min)$ or $(Max - Me)$ takes the place of the standard deviation.

These results show that only the balanced normalization proposed in this research simultaneously corrects the problems of uneven ranges, unequal minimum and maximum values, and asymmetry that imply an implicit weighting in the aggregation of variables (Table 4).

The longitudinal version of each normalization requires that the parameters be fixed over time. This is possible if the procedure is applied to stacked data, as the case study illustrates.

T tests identify significant changes for each variable in each quinquennium of the 2000 to 2020 period and for the whole period because the normalized data with the new procedure are comparable in time and space.

Table 4

Three desirable effects to control the implicit weighting of variables added in a composite index

Normalization procedure	Same maximum and minimum values for all variables	Same range for all variables	z-skewness and z-kurtosis in all variables are in the range ±2
Classic (z_{t0})	X	X	X
Normalized boxes (z_Q)	X	X	X
Min-Max (z_{MM})	✓	✓	X
Min-Max restricted (z_{MP})	X	✓	X
Robust normalization (z_{RN})	X[a)]	X[b)]	✓[c)]
Balanced normalization (z_{BN})	✓	✓	✓
Restricted balanced normalization (z_{RB})	✓	✓	✓

a) Same *Q1* and *Q3* values for all variables.
b) Same interquartile range (*Q1* to *Q3*).
c) Only in some variables.
Note: The author based on the methodology.
The methodology for quartiles *Q1* and *Q3* in z_{RN} is adapted to the lowest and highest values in z_{BS} and z_{RN}.

Conclusion

Without normalization, cross-sectional or longitudinal, there are no reliable composite indices. The selection of the appropriate normalization procedure is a current challenge and an ongoing process, as suggested in recent evaluations (Mazziotta–Pareto 2021, 2022; Sojda–Wolny 2020; Walesiak 2018). The comparative analysis of procedures shows that the key in the longitudinal extension of the cross-sectional versions is to maintain fixed parameters for stacked data, as follows:

a) Fixed mean and standard deviation (stdv) of a specific year. The mean and stdv of a reference year (e.g., 2000) are the same for all and every one of the raw values of the period (e.g., 2000 to 2020), as in z_{t0}. In Norman's version (2010, 2015, 2017), the mean and standard deviation are fixed because they correspond to all values of the period.

b) Minimum (min) and maximum (max) values of the fixed period, as in z_{MM}. This is equivalent to substituting in z_{t0} the mean for the minimum value of the period and the standard deviation for the min-max range of the period.
c) Average of a specific year (e.g., 2010) and fixed min and max values of the period, as in z_{MP}.
d) Median (Me) of the period and fixed min and max values, as in z_{BN}.

In variables of a composite index, it is not enough for the normalization to be longitudinal or absolute; it is also necessary to control the implicit weighting when adding them. This implicit weight is associated with the aggregation of variables with uneven ranges, unequal values at the ends, and asymmetry. z_{BN} is the only procedure that simultaneously corrects this triple problem.

In short, balanced normalization has several functions in a composite index (CI): a) to express the values in the same abstract unit that guarantees the cross-sectional and longitudinal comparison and aggregation of variables; b) to identify atypical cases; c) to invert the polarization of variables so that an increase in the variable is reflected in an increase in the IC; d) to allow statistical analysis and hypothesis-testing of information compatible in time and space; and d) to control the weighting implicit in i) the uneven ranges between the variables, ii) the unequal values at the extremes between variables, and iii) the asymmetry of each variable.

This work proposes and illustrates a normalization procedure; it does not suggest any explicit weighting procedure or variable aggregation method. These last two tasks deserve their own analysis. For example, it is necessary to consider the compensating or noncompensating nature of the composite index and measure the stability of the aggregation procedures by omitting one variable at a time. This exercise is the subject of other stages of research on composite indices.

Acknowledgments

This research is part of the general project "Spatial pattern of poverty in Mexico", code A1-S-40430, funded by the National Council of Science and Technology of Mexico (CONACYT).

The author appreciates Matteo Mazziotta and Adriano Pareto's kindness in sharing files, which helped him quickly understand their methodological developments in composite indices. Special thanks go to Paul Norman and an anonymous referee of the Regional Statistics journal for their comments and suggestions, which considerably improved the quality of the manuscript.

REFERENCES

ACHARYA, R.–PORWAL, A. (2020): A vulnerability index for the management of and response to the COVID-19 epidemic in India: An ecological study *The Lancet Global Health* 8 (9): e1142–e1151. https://doi.org/10.1016/s2214-109x(20)30300-4

ANGOA, I.–PÉREZ-MENDOZA, S.–POLÈSE, M. (2009): Los tres Méxicos: Análisis de la distribución espacial del empleo en la industria y los servicios superiores, por tamaño urbano y por región *EURE (Santiago)* 35 (104): 121–144. https://doi.org/10.4067/s0250-71612009000100006

BALCERZAK, A. P. (2016): Multiple-criteria evaluation of quality of human capital in the European Union countries *Economics and Sociology* 9 (2): 11–26. https://doi.org/10.14254/2071-789X.2016/9-2/1

BRIMICOMBE, A. (2000): Constructing and evaluating contextual indices using GIS: A case of primary school performance tables *Environment and Planning A: Economy and Space* 32 (11): 1909–1933. https://doi.org/10.1068/a3316

CREWS, K. A.–PERALVO, M. F. (2007): Segregation and fragmentation: Extending landscape ecology and pattern metrics analysis to spatial demography *Population Research and Policy Review* 27 (1): 65–88. https://doi.org/10.1007/s11113-007-9054-5

DASGUPTA, P.–WEALE, M. (1992): On measuring the quality of life *World Development* 20 (1): 119–131. https://doi.org/10.1016/0305-750x(92)90141-h

DĘBKOWSKA, K.– JAROCKA, M. (2013): The impact of the methods of the data normalization on the result of linear ordering *Acta Universitatis Lodziensis. Folia Oeconomica* (286): 181–188. http://hdl.handle.net/11089/10267

EXETER, D. J.–BOYLE, P. J.–NORMAN, P. (2011): Deprivation (im)mobility and cause-specific premature mortality in Scotland *Social Science & Medicine* 72 (3): 389–397. https://doi.org/10.1016/j.socscimed.2010.10.009

FLANAGAN, B. E.–GREGORY, E.W.–HALLISEY, E. J.–HEITGERD, J. L.–LEWIS, B. (2011): A social vulnerability index for disaster management *Journal of Homeland Security and Emergency Management* 8 (1). https://doi.org/10.2202/1547-7355.1792

GILTHORPE, M. S. (1995): The importance of normalisation in the construction of deprivation indices *Journal of Epidemiology & Community Health* 49 (Suppl 2): S45–S50. https://doi.org/10.1136/jech.49.suppl_2.s45

HAN, J.–PEI, J.–TONG, H. (2023): Data, measurements, and data preprocessing. In: *Data mining* pp. 23–84., Elsevier. https://doi.org/10.1016/B978-0-12-811760-6.00012-6

HEINRICH, K.–HUBER, C.–RINNER, C. (2017): Making area-based composite indices comparable across time – The role of standardization and outlier removal *URISA Journal* 27 (2): 37–49.

IL CHOI, H. (2019): Assessment of aggregation frameworks for composite indicators in measuring flood vulnerability to climate change *Scientific Reports* 9 (1): 19371. https://doi.org/10.1038/s41598-019-55994-y

KUKUŁA, K. (2014): Zero unitarisation method as a tool in ranking research *Economic Sciences for Rural Development* (36): 95–100.

LEYS, C.–LEY, C.–KLEIN, O.–BERNARD, P.–LICATA, L. (2013): Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median *Journal of Experimental Social Psychology* 49 (4): 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

ŁUCZAK, A.–JUST, M. (2020): A complex MCDM procedure for the assessment of economic development of units at different government levels *Mathematics* 8 (7): 1067. https://doi.org/10.3390/math8071067

ŁUCZAK, A.–JUST, M. (2021): Sustainable development of territorial units: MCDM approach with optimal tail selection *Ecological Modelling* 457: 109674. https://doi.org/10.1016/j.ecolmodel.2021.109674

MAZZIOTTA, M.–PARETO, A. (2021): Everything you always wanted to know about normalization (but were afraid to ask) *Rivista Italiana di Economia Demografia e Statistica* 65 (1): 41–52.

MAZZIOTTA, M.–PARETO, A. (2022): Normalization methods for spatio-temporal analysis of environmental performance: Revisiting the min–max method *Environmetrics* 33 (5): e2730. https://doi.org/10.1002/env.2730

NORMAN, P. (2010): Identifying change over time in small area socio-economic deprivation *Applied Spatial Analysis and Policy* 3 (2-3): 107–138. https://doi.org/10.1007/s12061-009-9036-6

NORMAN, P. (2015): The changing geography of deprivation in Britain: Exploiting small area census data 1971 to 2011. In: MALLESON, N.–ADDIS, N.–DURHAM, H.–HEPPENSTALL, A.–LOVELACE, R.– NORMAN, P.–OLDROYD, R. (eds.): *GIS Research UK (GISRUK) 2015 Proceedings* pp. 465–474., School of Geography, University of Leeds.

NORMAN, P.–DARLINGTON-POLLOCK, F. (2017): The Changing Geography of Deprivation in Great Britain: Exploiting Small Area Census Data, 1971 to 2011. Chapter 30. In: STILLWELL, J (ed.): *Routledge Handbook of Census Resources, Methods and Applications* pp. 404–420., Routledge, London.

RICKETTS HEIN, J.–ILBERY, B.–KNEAFSEY, M. (2006): Distribution of local food activity in England and Wales: An index of food relocalization *Regional Studies* 40 (3): 289–301. https://doi.org/10.1080/00343400600631533

RINNER, C.–PIETROPAOLO, A. (2021): Tracking area-based composite indices over time: change indicators vs. Jointly rescaled indicators *Geomatica* 75 (2): 51–64. https://doi.org/10.1139/geomat-2020-0014

SMITH, D. M. (1975): *Patterns in human geography: An introduction to numerical methods*. Douglas David & Charles limited, Canada.

SOJDA, A.–WOLNY, M. (2020): The impact of standardisation method on smart city ranking. Scientific Papers of Silesian University of Technology *Organization and Management Series* 142: 83–94. https://doi.org/10.29119/1641-3466.2020.142.6

WALESIAK, M. (2018): The choice of normalization method and rankings of the set of objects based on composite indicator values *Statistics in Transition New Series* 19 (4): 693–710. https://doi.org/10.21307/stattrans-2018-036

INTERNET SOURCES

CONEVAL Consejo Nacional de Evaluación de la Política de Desarrollo Social (2007): *Los mapas de Pobreza en México. Anexo técnico metodológico*. México. https://www.coneval.org.mx/rw/resource/coneval/med_pobreza/1024.pdf (downloaded: April 2023)

CONEVAL Consejo Nacional de Evaluación de la Política de Desarrollo Social (2021): *Índice de rezago social 2020 a nivel nacional, estatal, municipal y localidades.* https://www.coneval.org.mx/Medicion/IRS/Paginas/Indice_Rezago_Social_2020.aspx (downloaded: April 2023)

CUTILLO, A.–MAZZIOTTA, M.–PARETO, A. (2021): A composite index of economic well-being for the European Union countries. In: PERNA, C.–SALVATI, N.–SCHIRRIPA, F. (eds.): *Conference of the societá italiana di statistica (SIS) 2021* pp. 546–551., Pearson. https://it.pearson.com/content/dam/region-core/italy/pearson-italy/pdf/Docenti/Universit%C3%A0/pearson-sis-book-2021-parte-1.pdf (downloaded: April 2023)

IGLEWICZ, B.–HOAGLIN, D. C. (1993): *How to detect and handle outliers*. ASQ Quality Press. https://hwbdocuments.env.nm.gov/Los%20Alamos%20National%20Labs/TA%2054/11587.pdf (downloaded: April 2023)

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD) (2008): *Handbook on constructing composite indicators. Methodology and user guide* OECD Publications, Paris. https://www.oecd.org/sdd/42495745.pdf (downloaded: April 2023)

SIBLEY, D. (1987): *Spatial Applications of Exploratory Data Analysis* Geo Books, Norwich. https://alexsingleton.files.wordpress.com/2014/09/49-spatial-applications-of-exploratory-data-analysis.pdf (downloaded: April 2023)