

UNIVERSITÀ POLITECNICA DELLE MARCHE
Repository ISTITUZIONALE

A New Point of View in the Construction of Composite Indicators: A Simulation Illustration

This is the peer reviewd version of the followng article:

Original
A New Point of View in the Construction of Composite Indicators: A Simulation Illustration / Ciommi, M., Gigliarano, C., Mariani, F., Recchioni, M.C.. - STAMPA. - (2025), pp. 678-683. [10.1007/978-3-031-64350-7_114]

Availability:
This version is available at: 11566/366437 since: 2026-11-10T11:41:25Z

Publisher:
Springer

Published
DOI:10.1007/978-3-031-64350-7_114

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. The use of copyrighted works requires the consent of the rights' holder (author or publisher). Works made available under a Creative Commons license or a Publisher's custom-made license can be used according to the terms and conditions contained therein. See editor's website for further information and terms and conditions.
This item was downloaded from IRIS Università Politecnica delle Marche (https://iris.univpm.it). When citing, please refer to the published version.

Publisher copyright:

Springer (conference paper) - Postprint/Author's accepted Manuscript

This version of the conference paper has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: 10.1007/978-3-031-64350-7_114.

(Article begins on next page)

16 November 2026

A new point of view in the construction of composite indicators: a simulation illustration

Mariateresa Ciommi[1], Chiara Gigliarano[2], Francesca Mariani[1], and Maria Cristina Recchioni[1]

[1] Università Politecnica delle Marche, Piazzale Martelli 8, Ancona, Italy
m.ciommi@staff.univpm.itf.mariani@staff.univpm.it,
m.c.recchioni@staff.univpm.it
[2] Università LIUC, Corso Matteotti 22, Castellanza (Varese), Italy
cgigliarano@liuc.it

Abstract. The definition and construction of a composite indicator (CI) is a challenging operation. Here we propose a new technique to construct composite indicators that are based on the Box-Cox transformation. To illustrate our idea, a simulation study has been carried out. Results show that it is possible to find a parameter from the Box-Cox transformation that, in the least squares sense, ensures normality and, as a consequence, it opens new roads for researches in the field of the construction of CIs.

Keywords: composite indicators, box-cox, normal distribution

1 Introduction

In the last 20 years, the construction of composite indicators (CIs) has attracted an ever-increasing number of scholars. According to OECD [5], composite indicator is a mathematical aggregation of a set of individual indicators having, usually, no common unit of measurement and aiming at capturing a multidimensional phenomenon. However, its construction is a challenging operation that requires several steps and, among which the type of normalization and the aggregation technique.

Data transformation is one of open issues in CIs ([6]) that deserves to be in-depth investigated.

Here, we propose a new tool to construct CI based on the Box-Cox transformation [2] that maps non-normal variables into new variables approximately normally distributed. The use of a Box-Cox transformation is not new in literature. Berger and Casella [1] prove that the generalized means can be derived as least square estimates of transformed data. Mariani et al. [4] prove that in the p-transformed space, which is the space obtained by transforming a vector via the Box-Cox function of order p, the p-order generalized mean acts as the arithmetic mean.

Here, the transformed variables are then aggregated in order to get CIs and reliability on the values can be computed (Section (2)). To illustrate the idea, a simulation study is provided (Section (3)).

2 The method

2.1 Notation and basic definitions

Let $\mathbf{X}$ be a $n \times m$ matrix, in which n is the number of observations/units and m is the number of indicators. Let $\underline{\mathbf{X}}_j$ for $j = 1, \ldots, m$ be the $j-$th column of matrix $\mathbf{X}$ containing the values of the variable j, that is:

$$\mathbf{X} = [\underline{\mathbf{X}}_1, \underline{\mathbf{X}}_2, \ldots \underline{\mathbf{X}}_m],$$

where:

$$\underline{\mathbf{X}}_j = [x_{1j}, x_{2j}, \ldots, x_{mj}]^T, \quad j = 1, \ldots, m.$$

For $j = 1, 2, \ldots, m$, the data collected in the $j-$th column $\mathbf{X}_j$ is a sample from a random variable X_j, describing the $j-$th indicator.

2.2 Box-Cox transformation

The Box-Cox transformation [2] is a map which allows to transform non-normal observations into a normal shape. Given an $n \times 1$ vector of observations $\underline{x} = [x_1, x_2, \ldots, x_n]^T$ in which $x_i > 0$, $i = 1, 2, \ldots, n$, the Box–Cox transformation of parameter p, $p \in \mathbb{R}$, for $x_i > 0$ is defined as

$$h_p(x_i) = \begin{cases} \frac{x_i^p - 1}{p} & \text{if } p \neq 0, \\ \ln x_i & \text{if } p = 0. \end{cases} \tag{1}$$

Note that, for any p, the function $h_p(x)$, $x > 0$, is strictly increasing and satisfies the condition $h_p(1) = 0$.

Typically, the exponent p in (1) varies in the interval $[-5, 5]$.

As said above, the Box-Cox transformation requires all values in the input sequence to be positive and non-zero. However, it is possible to define the Box-Cox transformation also for negative values, by adding a second parameter $\lambda > 0$. Thus, the two-parameter Box–Cox transformations when $x_i > -\lambda$ is defined for $x_i > -\lambda$ as follows:

$$h_{p,\lambda}(x_i) = \begin{cases} \frac{(x_i + \lambda)^p - 1}{\lambda_1} & \text{if } p \neq 0, \\ \ln(x_i + \lambda) & \text{if } p = 0. \end{cases}$$

The *optimal value* for p is the one which results in the best approximation of a normal distribution curve, i.e. the value p for which the vector of the transformed observations $h_p(\mathbf{x}) = [h_p(x_1), h_p(x_2), \ldots, h_p(x_n)]^T)$ can be considered, approximately, as sampled from a variable normally distributed with constant variance.

Note that normality is an important assumption since it allows to run a broader number of tests.

The optimal value of p is typically determined by maximizing the log-likelihood function or minimizing another suitable criterion (for more details, we refer to [8] and [7]).

2.3 The model

For $j = 1, 2, \ldots, m$, let fix an indicator j, that is a column vector of data matrix $\mathbf{X}$ and we look for the value p such that the transformed vector according to equation (1), denoted as $\underline{\mathbf{Y}}_j$ is normally distributed.

Note that, the optimal values of p, associated with distincts indicators, are in general different.use of Box-Cox transformations lead to several values of p, one for each row of $\underline{\mathbf{X}}$. Thus, we denote the optimal value of p for indicator j as p_j, $j = 1, 2, \ldots, m$.

More in details, for $j = 1, 2, \ldots, m$, let $h_{p_j}(\mathbf{X}_j)$ be the Box-Cox transformation of parameter p_j of a vector $\mathbf{X}_j$, then

$$\mathbf{Y}_j = h_{p_j}(\mathbf{X}_j) = [h_p(x_{1j}), h_p(x_{2j}), \ldots, h_p(x_{nj})]^T \tag{2}$$

is, approximately, a sample drawn from a normal variable $Y_j \sim N(\mu_j, \sigma_j^2)$ with mean μ_j and variance σ_j^2.

Let us denote by $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y_2}, \ldots, \mathbf{Y}_m]$ the transformed matrix. Note that the $j-$th column of $\mathbf{Y}$ can be considered as a sample drawn from a normal distribution with mean μ_j and variance σ_j^2.

For $i = 1, 2, \ldots, n$, the $i-$th row of the transformed matrix $\mathbf{Y}$ contains a realization of the multivariate vector of normal variables $[Y_1, Y_2, \ldots, Y_m]^T$.

For each row of the matrix $\mathbf{Y}$ we compute the mean by columns:

$$z_i = \frac{1}{m} \sum_{j=1}^{m} h_{p_j}(x_{ij}), \quad i = 1, 2, \ldots, n,$$

and we collect these values in the vector $\mathbf{Z} = [z_1, z_2, \ldots, z_n]^T$.

It is worth noting that the vector $\mathbf{Z}$ contains the values of the random variable $Z \sim N\left(\frac{1}{m}\sum_{j=1}^{m}\mu_j, \frac{1}{m}\sum_{j=1}^{m}\sigma_j^2\right)$.

We define the new composite indicator as follows:

Definition 1 *Given the $n \times m$ data matrix $\mathbf{X} = \{x_{ij}\}_{i=1,2,\ldots,n, j=1,2,\ldots m}$, the composite indicator associated with the unit i is defined as:*

$$h_p^{-1}(z_i) = h_p^{-1}\left(\frac{1}{m}\sum_{j=1}^{m} h_{p_j}(x_{ij})\right), \quad i = 1, 2, \ldots, m,$$

where p is the solution of the following optimization problem:

$$\min_{p \in [-5,5]} F(p), \tag{3}$$

and:

$$F(p) = \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} h_p(x_{ij}) - z_i\right)^2. \tag{4}$$

We observe that, if, for each j, the value of $p_j = p$, that is the parameter is the same for the whole dataset, then, the Box-Cox transformation of order p of the dataset is the generalized mean of oder p (for more details, we refer to [3]):

$$h_p^{-1}(z_i) = h_p^{-1}\left(\frac{1}{m}\sum_{j=1}^{m} h_p(x_{ij})\right), \quad i = 1, 2, \ldots, m.$$

Note that the knowledge of the distribution of the random variable Z allows to use statistical inference techniques to attribute a reliability to the value of the indicators associated with different units. This could to make possible construct approximated confidence interval for the indicator values.

The existence and the uniqueness of the solution for problem (3), (4) is fundamental to define the composite indicator of Definition 1. For this reason, in the following section we investigate the behaviour of function $F(p)$ as p varies in a study case.

3 The simulation study

In order to show the feasibility of our methodology for constructing the new composite indicator defined in Definition 1, we solve the optimization problem (3), (4) in a numerical experiment using synthetic data.

The data are obtained sampling from a population of synthetic data. The population is made by $N = 10000$ units, for each of which $m = 5$ indicators are available. Each column is given by a sample of dimension N randomly generated from different distributions, namely *i)* Gamma distribution; *ii)* Inverse Gaussian distribution; *iii)* Exponential distribution; *iv)* Pareto distribution; and *v)* Weibull Distribution. Let $\mathbf{X}^{pop} = \{x_{i,j}^{pop}\}_{i=1,2,\ldots,N, j=1,2,\ldots,m}$ be the $N \times m$ matrix containing the values of the m indicators associated with the N units.

In Table 1 we show the parameters used to generate the synthetic sample.

Distribution	Parameter 1	Parameter 2
Gamma	shape=1	rate=0.8
Inverse Gaussian	mean=1.5	shape=0.5
Exponential	rate=0.3	
Pareto	shape=3	
Weibull	shape=1	rate=0.8

Table 1: Parameters of the distributions used to generate the data

To make the values of different indicators comparable with range in the interval $(0, 1]$, we normalize the data matrix $\mathbf{X}^{pop}$ by column using the the min-

max technique, i.e. we consider:

$$\tilde{x}_{ij}^{pop} = \frac{x_{ij}^{pop} - M_j^{pop}}{M_j^{pop} - m_j^{pop}}, \quad i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, m,$$

where $m_j^{pop} = \min(x_{1j}^{pop}, x_{2j}^{pop}, \ldots, x_{nj}^{pop})$ and $M_j^{pop} = \max(x_{1j}^{pop}, x_{2j}^{pop}, \ldots, x_{nj}^{pop})$ are, respectively, the minimum and the maximum values of the indicator j, for $j = 1, 2, \ldots, m$.

We extract a random sample of dimension $n = 1000$ and we collect the values of the random sample relative to the $m = 10$ indicators in a the 1000×5 data matrix $\mathbf{X}$. In order to assure the strictly positivity if data, needed to compute the Box-Cox transformation, we add to all data a constant equal to $1.e - 5$.

Then, according to our model, for each indicator j, $j = 1, 2, \ldots, m$, we find the value p_j that ensures that the transformed values $h_{p_j}(\mathbf{X}_j)$, $j = 1, 2, \ldots, m$, are normally distributed. As said in Section 2.2, in order to find the optimal value of p corresponding to the $j-$th indicator, i.e. p_j, we use the maximum likelihood estimation (for more details, we refer to [2]). Specifically, we use the *boxcox* function provided by the R package *MASS* (for more details, we refer to [9]).

The *boxcox* function computes the log-likelihood of data transformed through the Box-Cox power transformation associated with a vector of values of the parameter p in a given search grid, this is done adapting to the transformed data a linear regression model in the only intercept term is included. This allows to locate the transformation parameter under which the log-likelihood is maximized

Finally, we consider a grid of p values, $p_l = -5 + 0.01(l-1)$, $l = 1, 2, \ldots, 1001$, so the values considered are integer values varying from -5 to 5 and we evaluate the objective function $F(p)$ (see Eq. (4)) in this grid.

Figure 1) shows the function $F(p)$,, $i = 1, 2, \ldots$ of Problem (3), (4) when p varies from -5 to 5 (left panel) and from -1 to 1 (right panel).

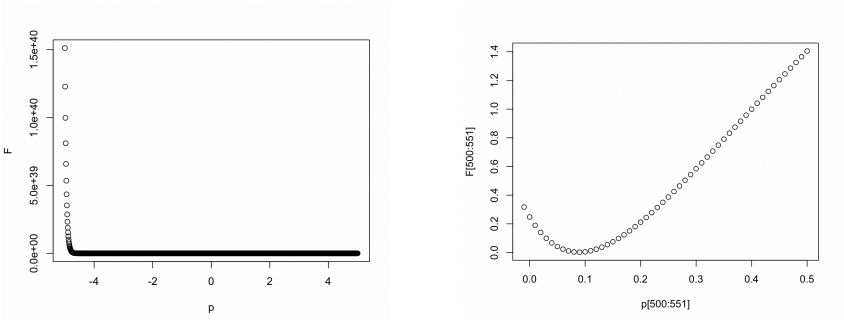


Fig. 1: Function $F(p)$ of problem (3), (4) when p varies from -5 to 5 (left panel) and from -0.011 to 0.5 (right panel).

Looking at the left panel of Figure 1 we can see that the objective function of problem (3), (4) has a burst for negative values of p, contrarily, for positive values of p the objective function is bounded between 0 and 1.5.

Looking at the right panel of Figure 1 and computing the minimum value in the set $\{F(p_l) : p_l = -5 + 0.01(l-1),\ l = 1, 2, \ldots, 1001\}$ we find that the function $F(p)$ attains its minimum value at $p = 0.09$.

4 Conclusions

Here, we illustrate a new proposal for construction composite indicators by transforming each variable by means of a Box-Cox transformation of order p. By definition of Box-Cox transformation, the new variables are approximately normally distributed. Thus, we can associate to the corresponding CIs a reliability factor aiming at discriminating two consecutive units. This work is only a first step in the direction of building CIs with a reliability factor on the corresponding ranking.

References

1. Berger, R.L.; Casella, G.: Deriving Generalized Means as Least Squares and Maximum Likelihood Estimates, The American Statistician, 46(4), 279-282 (1992).
2. Box, G. E. P., Cox, D. R.: An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, 211-252 (1964). https://doi.org/10.1111/j.2517-6161.1964.tb00553.x
3. Bullen, P. S.: Handbook of Means and Their Inequalities. Dordrecht, Netherlands: Kluwer, 175-177 (2003). https://link.springer.com/book/10.1007/978-94-017-0399-4
4. Mariani, F., Ciommi, M., Recchioni, M. C.: A new class of composite indicators: the penalized power means (2022). arXiv preprint arXiv:2206.11216.
5. OECD/European Union/EC-JRC:, Handbook on Constructing Composite Indicators: Methodology and User Guide, OECD Publishing, Paris (2008). doi: 10.1787/9789264043466-en
6. Pareto, A., Grimaccia, E., Mazziotta, M., Terzi, S.: Open issues in composite indicators. A starting point and a reference on some state-of-the-art issues 2021. Roma TrE-Press. https://romatrepress.uniroma3.it/wp-content/uploads/2021/03/open-togmp-2.pdf
7. Sakia, R. M.: The Box-Cox transformation technique: a review. Journal of the Royal Statistical Society Series D: The Statistician, 41(2), 169-178 (1992).
8. Spitzer, J. J.: A primer on Box-Cox estimation. The Review of Economics and Statistics, 307-313 (1982).
9. Venables, W.N., Ripley, B.D. : Modern Applied Statistics with S. Fourth edition. Springer, New York. ISBN 0-387-95457-0 (2002). http://www.stats.ox.ac.uk/pub/MASS4