Optimising Composite Indicators with Sensitivity Analysis

William Becker[1], Michaela Saisana[1], Paolo Paruolo[1], Ine Vandecasteele[2] and Andrea Saltelli[3,4]
[1] European Commission, Joint Research Centre, Deputy Director-General, Econometrics and Applied Statistics Unit, Via E Fermi 2749, Ispra (VA), Italy
[2] European Commission, Joint Research Centre, Institute for Environment and Sustainability, Sustainability Assessment Unit, Via E Fermi 2749, Ispra (VA), Italy
[3]Centre for the Study of the Sciences and the Humanities (SVT), University of Bergen (UIB), Spain
[4]Institut de Ciència i Tecnologia Ambientals (ICTA), Universitat Autonoma de Barcelona, Spain
Email: william.becker@jrc.ec.europa.eu

Multi-dimensional measures (often termed composite indicators) are popular tools in the public discourse for assessing the performance of countries/entities on human development, perceived corruption, innovation, competitiveness, or other complex phenomena that are not directly measurable and not precisely defined. These measures combine a set of relevant variables using an aggregation formula, which is often a weighted arithmetic average. The values of the weights are usually meant to reflect the variables' importance in the index, which is based on the subjective beliefs of the developer. In practice, however, correlations between variables mean that the weights assigned to each variable do not actually reflect the true importance in terms of their contribution to the composite indicator. To elaborate, let $\{x_i\}_{i=1}^{d}$ be the set of *d* input variables to the composite indicator, and *y* be the output (i.e. the composite indicator value). Weights w_i are assigned such that:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d,$$

where $\sum_{i=1}^{d} w_i = 1$. Given a sample of *N* points, consider an importance measure *I* which measures the influence of each x_i on y, which is also normalised to sum to 1. The key point is that $I_i \neq w_i$, nor is *I* necessarily linearly related to *w*, although this fact is sometimes overlooked by developers. Note that the importance of a composite indicator's inputs on its outputs is dependent on the sample. Two questions immediately arise: first, given a set of weights and a sample, what is the influence of each variable on the output? Second, how can weights be assigned to reflect the desired importance?

This work views the problem from a sensitivity analysis perspective, using tools from the literature on global sensitivity analysis with correlated inputs to understand the importance of each variable's contribution to the index. In particular, the first order sensitivity index, $S_i = \mathrm{var}[\mathrm{E}(Y|X_i)]/\mathrm{var}(Y)$ is used, which is referred to here as the *Pearson correlation ratio* (an equivalent term that was used by Karl Pearson first in 1905). This term is used to emphasise that the measure is being used first and foremost to measure correlation, and not sensitivity in terms of a variance decomposition.

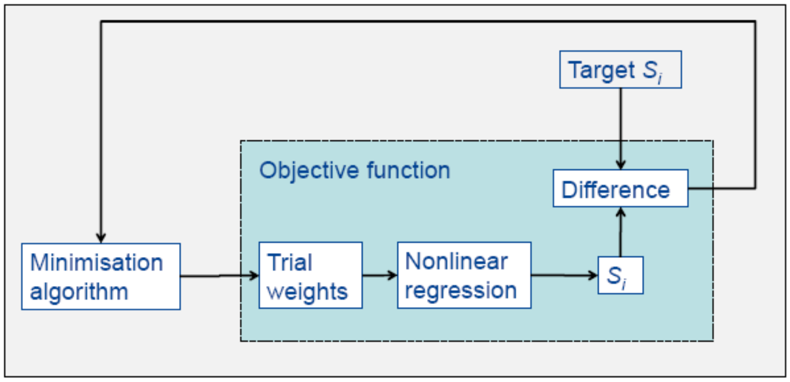
In order to estimate the correlation ratio, this work follows previous work of [3] and [1] by using nonlinear regression to estimate the main effect $\mathrm{E}(Y|X_i)$. However, additional to the use of local polynomial regression, two other approaces are considered. The first is the use of *penalised splines*, which can be fit with a particularly low computational cost (an advantage which is exploited in the optimisation step below). The second is the use of Bayesian *Gaussian processes*, which have the advantage of providing confidence intervals on the Pearson correlation ratio.

As a further step, to better understand the influence of variables on the composite indicator, an approach based on the correlated sensitivity analysis work of [4] is applied, which uses an additional regression to decompose the influence of each variable into influence caused by correlation, and influence caused by the composite indicator structure (aggregation and weights).

To now address the second question, the issue of optimisation of weights is considered. Although this problem has been tackled in [3] using linear regression, the proposal here is to extend it to nonlinear regression, to account for nonlinear main effects. Letting $\tilde{S}_i$ be the desired correlation ratio of variable x_i, the set of weights $\boldsymbol{w}_{\text{opt}}$ that minimises the difference between $\tilde{S}_i$ and $S_i(\boldsymbol{w})$ is found by,

$$\boldsymbol{w}_{\text{opt}} = \operatorname{argmin} \sum_{i=1}^{d} \left(\tilde{S}_i - S_i(\boldsymbol{w})\right),$$

where $\boldsymbol{w} = \{w_i\}_{i=1}^{d}$. This minimisation problem is performed by the Nelder-Mead simplex search method [2]. See the figure below for an overview of the optimisation process.



The methodologies proposed here are applied to several test cases, namely the Resource Governance Index, the Good Country index, and a hydrological example—the Water Retention index, which demonstrates how weight optimisation can be performed on composite indicators with thousands, or possibly millions, of data points. The case studies provide insight in terms of the ideal weightings for each composite indicator, as well as illustrating the potential and limitations of the proposed approaches.

[1] Da Veiga, S., Wahl, F., & Gamboa, F. (2009). Local polynomial estimation for sensitivity analysis on models with correlated inputs. *Technometrics*, *51*(4), 452-463.

[2] Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder--Mead simplex method in low dimensions. *SIAM Journal on optimization*, *9*(1), 112-147.

[3] Paruolo, P., Saisana, M., & Saltelli, A. (2013). Ratings and rankings: voodoo or science?. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(3), 609-634.

[4] Xu, C., & Gertner, G. Z. (2008). Uncertainty and sensitivity analysis for models with correlated parameters. *Reliability Engineering & System Safety*, *93*(10), 1563-1573.