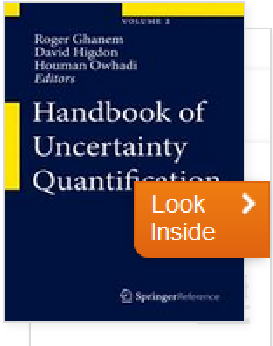
Reference Work Entry
Handbook of Uncertainty Quantification
pp 1187-1216

Date: 17 June 2017

Weights and Importance in Composite Indicators: Mind the Gap

William Becker , Paolo Paruolo, Michaela Saisana, Andrea Saltelli



Abstract

Multidimensional measures (often termed composite indicators) are popular tools in the public discourse for assessing the performance of countries on human development, perceived corruption, innovation, competitiveness, or other complex phenomena. These measures combine a set of 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. This paper uses measures drawn from global sensitivity analysis, specifically the Pearson correlation ratio, to discuss to what extent the importance of each variable coincides with the intentions of the developers. Two nonparametric regression approaches are used to provide alternative estimates of the correlation ratios, which are compared with linear measures. The relative advantages of different estimation procedures are discussed. Three case studies are investigated: the Resource Governance Index, the Good Country Index, and the Financial Secrecy Index.

Keywords

Composite indicators – Nonlinear regression – Sensitivity analysis

Weights and Importance in Composite Indicators: Mind the Gap

Chapter · March 2016

DOI: 10.1007/978-3-319-11259-6_40-1

4 authors:

William Becker
European Commission

Paruolo Paolo
European Commission

Michaela Saisana
European Commission

Andrea Saltelli
University of Bergen

Cite this entry as:

Becker W., Paruolo P., Saisana M., Saltelli A. (2017) Weights and Importance in Composite Indicators: Mind the Gap. In: Ghanem R., Higdon D., Owhadi H. (eds) Handbook of Uncertainty Quantification. Springer.

Title: Weights and Importance in Composite Indicators: Mind the Gap

Name: William Becker[1], Paolo Paruolo[1], Michaela Saisana[1], Andrea Saltelli[23]

Affil./Addr. 1: European Commission, Joint Research Centre

Via Enrico Fermi, 2749

21027 Ispra VA, Italy

email: william.becker, paolo.paruolo, michaela.saisana@jrc.ec.europa.eu

Affil./Addr. 2: Centre for the Study of the Sciences and the Humanities (SVT)

University of Bergen (UIB), Norway

Affil./Addr. 3: Institut de Ciencia i Tecnologia Ambientals (ICTA)

Universitat Autonoma de Barcelona (UAB), Spain

E-mail: andrea.saltelli@svt.uib.no

Weights and Importance in Composite Indicators: Mind the Gap

Summary. Multi-dimensional measures (often termed composite indicators) are popular tools in the public discourse for assessing the performance of countries on human development, perceived corruption, innovation, competitiveness, or other complex phenomena. These measures combine a set of 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. This paper uses measures drawn from global sensitivity analysis, specifically the Pearson correlation ratio, to discuss to what extent the importance of each variable coincides with the intentions of the developers. Two nonparametric regression approaches are used to provide alternative estimates of the correlation ratios, which are compared with linear measures. The relative advantages of different estimation procedures are discussed. Three case studies are investigated: the Resource Governance Index, the Good Country Index, and the Financial Secrecy Index.

Introduction

Composite indicators are aggregations of observable variables (indicators) that aim to quantify underlying concepts that are not directly observable, such as competitiveness, freedom of press or climate hazards. Composite indicators are employed for many purposes, including policy monitoring, and they are called in several different ways; for instance they are also referred to as "performance indices".

Hardly any newspaper can resist the temptation of commenting on an international country ranking. The popularity of rankings owes to two main reasons. First, their simplicity. They provide a summary picture of the multiple facets or dimensions of complex, multidimensional phenomena in a way that facilitates evaluation and comparison. Second, rankings force institutions and governments to question their standards; rankings are drivers of behaviour and of change (Kelley and Simmons, 2015).

Hence, it comes at no surprise that there has been a turbulent growth of performance indices over the past two decades. Bandura (2011) provides a comprehensive inventory of over 400 country-level indexes monitoring complex phenomena from economic progress to educational quality. Similarly, a more recent inventory by the United Nations (Yang, 2014) details 101 composite measures of human well-being and progress, covering a broad range of themes from happiness-adjusted income to environmentally-adjusted income, from child development to information and communication technology development. Several of those indices have been cited online more than a million times.

The construction of a composite indicator requires several choices; it involves a number of steps in which the developer must make decisions regarding which variables to include in the composite index, and how to aggregate them. These steps involve first, developing a conceptual framework, then selecting and treating data sets. Next, a multivariate analysis is often performed to identify principal components and correlations. The data are then normalised, for example by scaling onto the unit interval or adjusting by mean and variance.

The step discussed in this chapter is the aggregation step, where typically the variables are combined in a weighted average to give the resulting value of the composite indicator. Apart from the decision of which kind of weighted average to use (e.g. arithmetic, geometric), the developer must select values of weights to apply to each variable. The values of these weights can have a large impact on the subsequent rankings of the composite indicator. Understanding the impact of weights on the output composite indicator is hence important.

A possible misconception is that the weight assigned to a variable can be directly interpreted as a measure of importance of the variable to the resulting value of the composite indicator. However this is true only under very restrictive assumptions; different variances and correlations among variables, for instance, prevent the weights from corresponding to the variables' importance.

A common practice in composite indicator construction is to set the weights of each input variable to be equal, with the intention to make each input variable contribute equally to the value of the composite indicator. This is usually done because choosing weights other than equal would impose questionable assumptions on the relative importance of each variable. However, as shown in this chapter, nominal weights rarely coincide with variable importance, because of dependence between input variables, or – when the variables have not been standardized to have the same variance, simply because of variance differences among inputs.

In this chapter, tools from sensitivity analysis are applied to address the question: how dependent is the composite indicator, considered as an *output* variable, with respect to each single measured variable (the *input* variable) which is used to build it? This question concerns the relative importance of input variables in the composite indicator, where the terms *input* and *output* follow the terminology used in uncertainty and sensitivity analysis, the wider subject of this book.

This chapter highlights how the relative importance of input variables should not be confused with the nominal weights that are used to construct the composite indicator. Other aspects that influence the issue of the importance of input variables are the dependence between variables and the possible nonlinear transformations applied to input variables in the aggregation. The chapter reviews an earlier proposal of some of the present authors, see Paruolo et al (2013), to measure the relative importance with the Pearson correlation ratio between the composite indicator and the input variables. Indeed, this measure of importance differs from the nominal weights used in the aggregation.

The chosen measure of importance in this chapter is the Pearson's *correlation ratio*, which is a variance-based measure that accounts for (possibly nonlinear) dependence between input variables and the output. It is exactly equivalent to the *main effect index* or *first order sensitivity index* (as it is more widely referred to), but is termed the correlation ratio here, first to avoid confusion of the term "index" with composite indices, and second to emphasise that it is used as a measure of nonlinear dependence, rather than uncertainty. This distinction is explained in a little more detail later.

According to the authors' experience, the fact that weights are not measures of importance is rather counter intuitive at first sight. The proof of this fact is that most developers do indeed use weights with the intention to tune the importance of variables in the composite index. As shown in Paruolo et al (2013), a simple glance at the scatterplots of the composite indicator versus its input variables is sufficient to convince the reader that the relative importance of variables may be quite far from what the weights would imply. Denote the output of a composite indicator as y, which is a function of several input variables $x_1, x_2, ...$ Now, consider an example in which inputs x_1 and x_2 are correlated with each other, but both independent from a third input x_3. In this case, the scatterplots of y against each input variable would show (qualitatively) that—even if the three variables have equal weights and variances—the importance of x_1 and x_2 (in terms of the effect on y) would be larger than that of x_3 due to the dependence between x_1 and x_2.

This example is important because in general, the variables in a composite indicator are correlated; they need to be, as one assumes that they concur to describe a unique latent phenomenon. In spite of this rule, the reader is now asked to entertain the following thought experiment, where a composite indicator is built using just two uncorrelated variables (or pillars) with the same variance. The purpose of the experiment is to show the counter-intuitiveness of the relation between weights and importance. As a rule, weights in a composite indicator are set by their developers to add up to one. Thus if one wishes to have, for example, variable x_1 more important than x_2 one could assign weight $w_1 = 0.9$ to x_1 and $w_2 = 0.1$ to the x_2. This would imply that x_1 drives y much more than x_2.

Would this translate into a quantitative statement such as "x_1 accounts for 90% of the variance of y, while x_2 accounts for 10%"? No, because if one decomposes the variance of y according its uncorrelated inputs, the fractional importance of a given variable x_i (i.e. the correlation ratio just mentioned) is $w_i^2 / \sum_i w_i^2$ for the case where all variables have the same variance, (this is an exercise in the sensitivity analysis handbook Saltelli et al (2008)). In the case above, this implies fractional importance measures of x_1 and x_2 equal to 99% and 1% respectively. Thus in order for the weights to add up to one, and have target importance 90% and 10%, weights should in fact be equal to 3/4 and 1/4, respectively.

Further justification for using the correlation ratio as the right importance measure to use for variables importance is provided in the remainder of this chapter, where the idea of target fractional variance to be achieved is discussed by judicious assignment of weights for the realistic case where the variables are not independent.

The correlation ratio can be estimated non-invasively (i.e. with only a set of input values and corresponding output values, and no explicit modelling of uncertainties as in (Saisana et al, 2005)), using relatively simple, nonlinear regression tools; Paruolo et al (2013) proposed to use non-parametric, local linear, kernel regression. This method is similar to the one in Da Veiga et al (2009), but does not require to use an independent sample from the marginal distribution of x to estimate the variance of the conditional expectation. In this chapter, the use of penalised splines is also reviewed in this context, see Ruppert et al (2003). The relative merits of different estimation methods for the conditional expectations are discussed. Results obtained for nonlinear regression are compared to with those obtained by linear regression, and the added value of nonlinearity is assessed in concrete cases studies.

The approach of this chapter lends itself to the possibility of selecting nominal weights which imply the intended importance (correlation ratios) of each input variable by searching through the "weight-space"(for a short discussion on reverse engineering the weights see (Paruolo et al, 2013)). An additional feature of this approach is that, as by-product of the analysis, it provides an estimate of the conditional expectation of the composite indicator as a function of a single input variable. The local slope of this conditional expectation answers the related research question: 'how much would the composite indicator increase for a marginal increase of the input variable (averaged over variations in other variables)?'. This question may be of interest when discussing alternative policy measures geared to influence different input variables.

The remainder of this chapter is organised as follows: the construction of composite indicators and some measures of variables' importance are reviewed first, including the correlation ratio. A description is then given of linear and nonlinear approaches to estimation of the correlation ratio. Three case studies are used to illustrate the relative merits of each estimation approach: the Resource Governance Index, which aims to measure transparency and accountability in the oil, gas and mining sectors; the Financial Secrecy Index, which measures secrecy and scope for abuse in the financial sector for each country; and the Good Country Index, which aims to measure to what extent a given country contributes to the some preestablished normative 'goods' for humanity. In these case studies, the relative strengths of the correlation ratio compared to linear measures of dependence are also discussed.

Measures of Importance and Transformations

Transformations and Weighting

Consider the case of a composite indicator y (output) calculated aggregating over d input variables x_i. The most common aggregation scheme is the weighted arithmetic average, i.e.

$$y_j = \sum_{i=1}^{d} w_i x_{ji}, \qquad j = 1, 2, \cdots, n \tag{1}$$

where x_{ji} is the normalised score of individual j (e.g., country) based on the value X_{ji} of the ith raw variable $i = 1, \ldots, d$ and w_i is the nominal weight assigned to the i-th variable X_i or x_i. Input variables are usually normalised according to the min-max normalisation method, see Bandura (2011),

$$x_{ji} = \frac{X_{ji} - X_{\min,i}}{X_{\max,i} - X_{\min,i}}, \tag{2}$$

where $X_{\max,i}$ and $X_{\min,i}$ are the upper and lower values respectively for the variable X_i; in this case all scores x_{ji} vary in $[0, 1]$. $X_{\max,i}$ and $X_{\min,i}$ could be replaced by maximal and minimal values for the X_i that do not depend on the sample observations.

A popular alternative to the min-max normalisation in (2) is given by the standardisation

$$x_{ji} = \frac{X_{ji} - \mathrm{E}(X_i)}{\sqrt{\mathrm{V}(X_i)}}, \tag{3}$$

where $\mathrm{E}(X_i)$ and $\mathrm{V}(X_i)$ are the mean and variances of the original variables X_i. Here $\mathrm{E}(X_i)$ and $\mathrm{V}(X_i)$ can be estimated by the sample mean and variance. Note that (3) guarantees equal variances, while transformation (2) does not.

In fact, transformation (2) scales each input variable onto $[0, 1]$, but allows for different means and variances. On the other hand, (3) transforms all variables such that they have a mean of zero and a variance of one. Importantly however, both transformations do not affect the correlation structure of variables, because both are linear transformations of the original variables and correlations are invariant with respect to linear transformations.

The weight, w_i, attached to each variable, x_i, is usually chosen so as to reflect the importance of that variable with respect to the concept being measured. The ratios w_i/w_ℓ can be taken to be the "revealed target relative importance" of inputs i and ℓ because they measure the substitution effect between x_i and x_ℓ, i.e. how much x_ℓ must be increased to offset or balance a unit decrease in x_i, see Decancq and Lugo (2013).

One of the central messages in this chapter is that the target importance does not usually coincide with the “true” importance of each variable, as defined by the correlation ratio, as shown above for the trivial case of two uncorrelated variables. In order to better understand the contribution of each input variable to the output of the composite indicator, measures of linear and nonlinear dependence may be used; these are reviewed in the following section.

Importance measures

In the following, the sample index subscript j (usually representing the country or region) is dropped unless it is needed for clarity. Let also the expected value and variance of y be $\mathrm{E}(y) = \mu_y$ and $\mathrm{V}(y) = \sigma_y^2$. Similarly, the means and variances of each x_i are denoted as $\{\mu_i\}_{i=1}^d$ and $\{\sigma_i^2\}_{i=1}^d$ respectively.

For any given composite indicator, one can define measures of importance of each of the input variables with respect to the output values of the composite indicator. One approach for assessing the influence of each input variable x_i on the composite indicator is to measure the dependence of y on x_i, where from here on, variables are assumed to be normalised – see equations (2) and (3).

Assume that

$$y_j = f_i(x_{j,i}) + \varepsilon_j \tag{4}$$

where $f_i(x_{j,i})$ is an appropriate function, possibly nonlinear, that models $\mathrm{E}(y_j|x_i)$ – the conditional expectation of y given x_i – and ε_j is an error term. Dependence of y on x_i can be measured in a number of ways. The covariance and correlation between x_i and y, for example, are defined as

$$\mathrm{cov}(y, x_i) := \mathrm{E}[(y - \mu_y)(x_i - \mu_i)], \qquad R_i := \mathrm{corr}(y, x_i) := \frac{\mathrm{cov}(y, x_i)}{\sigma_y \sigma_i}. \tag{5}$$

Remark here that $\mathrm{corr}(y, x_i)$ is a standardised version of the covariance, called the *Pearson product-moment correlation coefficient*, which scales the covariance onto the interval $[-1, 1]$. In the case of a simple linear regression of y on x_i, the square of the correlation coefficient gives the well-known linear *coefficient of determination* R^2, i.e. $R_i^2 := \mathrm{corr}^2(y, x_i)$, which takes values in $[0, 1]$.

The coefficient of determination is used to measure the goodness-of-fit of an ordinary linear regression: as such R_i^2 is a measure of linear dependence. Because of (5), the covariance $\mathrm{cov}(y, x_i)$, the correlation $\mathrm{corr}(y, x_i)$, and the coefficient of determination R_i^2 are all related measures of linear dependence. In sample, the coefficient of determination R_i^2 can be computed as

$$SS_{\mathrm{reg},i}/SS_{\mathrm{tot}}, \tag{6}$$

where $SS_{\text{reg},i} = \sum_{j=1}^{n}(\widehat{f}_i(x_{i,j}) - \bar{y})^2$ is the sum of squares explained by the linear regression, $\bar{y} := n^{-1}\sum_{j=1}^{n} y_j$ is the sample average, $\widehat{f}_i(x_{i,j}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,j}$ is the linear fit for observation y_j and $SS_{\text{tot}} = \sum_j (y_j - \bar{y})^2$ is the total sum of squares. R_i^2 can hence be seen as the ratio of the sum of squares explained by the linear regression of y on x_i, and the total sum of squares of y.

If the relationship between y on x_i is nonlinear, R_i^2 may underestimate the degree of dependence. The proposed measure in this chapter is the *correlation ratio*, S_i, $i = 1, 2, ..., d$ also widely known as the *first order sensitivity index*, or *main effect index* (see Chapter on Variance-based Sensitivity Analysis: Theory and Estimation Algorithms). This measure is meant to measure the (possibly nonlinear) influence of each variable on the composite indicator. The correlation ratio can be interpreted as the expected variance reduction of the composite indicator, if a given variable were fixed. The correlation ratio is traditionally denoted as η_i^2 and was introduced by Pearson (1905); to follow the wider literature on sensitivity analysis it is referred to here as S_i. Both the correlation ratio and the first order sensitivity index are defined as:

$$S_i \equiv \eta_i^2 := \frac{\mathrm{V}_{x_i}\left(\mathrm{E}_{\mathbf{x}_{\sim i}}\left(y \mid x_i\right)\right)}{\mathrm{V}\left(y\right)}, \tag{7}$$

where $\mathbf{x}_{\sim i}$ is defined as the vector containing all the variables $(x_1, \ldots, x_d)$ except variable x_i and $\mathrm{E}_{\mathbf{x}_{\sim i}}\left(y \mid x_i\right)$ denotes the conditional expectation of y given x_i, e.g. with x_i fixed at one value in its interval of variation. The notation employed here stresses that the expectation in $\mathrm{E}_{\mathbf{x}_{\sim i}}\left(y \mid x_i\right)$ is computed with respect to the distribution of $\mathbf{x}_{\sim i}$, i.e. with respect to all other input variables; while the subscript x_i used the outer variance signifies that the variance is taken over all possible values of x_i. In the following the variance and expected value subscripts are dropped to economise on notation.

The conditional expectation $\mathrm{E}\left(y \mid x_i\right)$ is known as the *main effect* of x_i on y, and it describes the expected value of y (the composite indicator) averaged over all input variables except x_i, keeping x_i fixed. As such, $\mathrm{E}\left(y \mid x_i\right)$ is a function of x_i and is here denoted as $f_i(x_i)$. This function is not typically known, however it can be estimated by performing a (nonlinear) regression of y on x_i. Various approaches for this problem are discussed in the following section; any of them delivers a fitted value $m_j := \widehat{f}_i(x_{i,j})$ corresponding to the prediction of y_j. The correlation ratio S_i can then be computed in sample as

$$\sum_{j=1}^{n}(m_j - \bar{m})^2 / \sum_{j=1}^{n}(y_j - \bar{y})^2 \tag{8}$$

where $\bar{m} := n^{-1}\sum_{j=1}^{n} m_j$, $m_j := \hat{m}(x_{ji})$ and $\hat{m}(\cdot)$ is the estimate of $m(\cdot) := f_i(\cdot)$. Eq. (8) mimics (6).

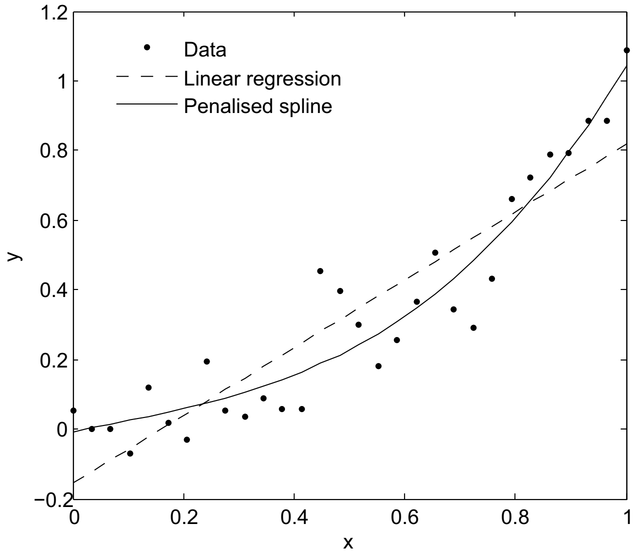


Fig. 1: Parabolic conditional mean with linear and penalised spline fits. Here $\operatorname{corr}(x_1, y) = 0.91$, $R_1^2 = 0.82$, and $S_1 = 0.91$ (using the penalised spline estimate).

The correlation ratio S_i is closely connected with the measure R_i^2 of linear dependence discussed previously. More specifically, R_i^2 equals S_i when $f_i(x_i)$ is linear: note that the main effect is dependent on the functional form of the composite indicator *and* the dependence structure between variables. This shows that, in the linear case, S_i is also related to the covariance and correlation coefficients by the associations discussed earlier.

In order to illustrate why linear measures of importance would not be sufficient for nonlinear dependence, Figure 1 plots data generated using the relationship $y = 0.2 + x_1^2 + 0.1u$, with $u \sim \mathcal{N}(0, 1)$ independent of x_1, which is generated drawing from a uniform distribution on $[0, 1]$. The conditional mean $\mathrm{E}(y \mid x_1) = x_1^2$ is here nonlinear (a branch of a parabola). The linear regression through the data in Figure 1 gives $R^2 = 0.82$, while the correlation ratio S_i is is 0.91, as estimated by a nonlinear (penalised spline) regression.

In this case the strong dependency between y and x_1 would be underestimated by the linear measure R_i^2, but it is captured by the correlation ratio S_i: about 91% of the variance in y is explained by the correlation ratio (using a nonlinear conditional mean specification), and about 9% of it would be missed by a linear regression model.

Relation to sensitivity analysis

The correlation ratio is widely used in the discipline of variance-based sensitivity analysis of computer models, typically under the name of "first order sensitivity index", "main effect index" or "Sobol' index". It is important to recognise that sensitivity analysis, as it is commonly performed in physical sciences and engineering, is concerned with analysing the effects of *uncertainty* in model input variables on the model output. Each input variable therefore has an associated probability distribution $p(x_i)$ which attempts to characterise the uncertainty in the value of x_i, either as a result of lack of knowledge or inherent variability. The "sensitivity analysis" described in this chapter, on the other hand, does not deal with uncertainty in the inputs of a composite indicator (although this is also an important area of research—see e.g. Saisana et al (2011)). Instead, it quantifies the contribution of each input to the composite indicator, as a result of the weight and aggregation of the indicator and the distribution and inter-dependence of each input variable, by measuring the nonlinear dependence of the composite indicator on each of its input variables. The distributions $p(x_i)$ of each input variable do not therefore characterise uncertainty in x_i—rather, they simply represent the distribution of entities (e.g. countries, institutions) in the variable x_i. Indeed, these distributions are often discrete: they represent a finite number of entities, such as countries or regions. It is therefore to emphasise that the application here is distinct from uncertainty analysis that the term "correlation ratio" is used, as opposed to "first order/main effect index", as more widely used elsewhere.

Usually in sensitivity analysis, as it is applied in physical models in science and engineering, one finds $\sum_{i=1}^{d} S_i \leq 1$ when studying a generic nonlinear output function with independent input variables, see Saltelli et al (2000); Li et al (2010). In this setting, the variance of the model output can be decomposed into portions that can be uniquely attributed to each variable and subset of variables. In the case of composite indicators, the input variables are almost certainly (positively) correlated, and it is usually the case that $\sum_{i=1}^{d} S_i > 1$. Furthermore, the correlation ratios here are not interpreted in terms of a variance decomposition—they are simply used as nonlinear measures of correlation. However, even for the case of non-independent inputs, the S_i measure preserves its meaning of expected fractional reduction of the output variance that would be achieved if a variable could be fixed (Saltelli and Tarantola, 2002).

In order to estimate $f_i(x_i)$, a wide number of approaches are available. Two nonlinear regression approaches are discussed in the next section; both offer a flexible framework for estimating nonlinear main effects.

Estimating Main Effects

This section reviews the estimation of the conditional expectation $\mathrm{E}(y \mid x_i)$ using polynomial (linear) regression, penalised splines and non-parametric regression. There are many other ways of modelling $\mathrm{E}(y \mid x_i)$, such as Gaussian processes (Rasmussen and Williams, 2006); see also the chapter on Sampling-based Procedures for Uncertainty and Sensitivity Analysis in this book, and Storlie and Helton (2008). The methods reviewed here are chosen for comparison purposes and because they have some attractive properties. Specifically, the linear regression model is considered here as the simplest model of the conditional expectation; kernel non-parametric regression is considered as the opposite polar leading example of nonlinear models for the conditional expectations; this method was employed in Paruolo et al (2013) in the context of composite indicators. As an additional method to estimate the nonlinear regression function, penalised splines are also investigated here, which share similar properties in smoothing nonlinear data with kernel non-parametric regression; moreover, they are computationally fast (and so can cope with large datasets). Note that in this section the variable subscript i is dropped, since the techniques here all refer to regression of y on a single variable x.

Polynomial Regression

Consider a set of bivariate data consisting of n pairs of observations $\{x_j, y_j\}$, $j = 1, 2, ..., n$, a integer p (the degree of the polynomial) and the model

$$y_j = \beta_0 + \beta_1 x_j + \cdots + \beta_p x_j^p + \epsilon_j \qquad j = 1, 2, ..., n \tag{9}$$

in which the β_h, $h = 0, \ldots, p$ are coefficients and ϵ_j is an error term. This can be rewritten in matrix form using the $p+1$ column vectors $\mathbf{x}_j = (1, x_j, x_j^2, ..., x_j^p)^\top$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ and the n-dimensional column vectors $\mathbf{y} = (y_1, y_2, ..., y_n)^\top$ and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^\top$. Next, one can define the $n \times (p+1)$ design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ so as to rewrite (9) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{10}$$

Minimising the sum of squared residuals with respect to $\boldsymbol{\beta}$ gives the well-known expression

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{11}$$

In case $p = 1$, (11) gives the linear regression coefficients $\widehat{\beta}_0, \widehat{\beta}_1$, which can be used to calculate R^2. Note that in cases where $p > 1$, (10) is nonlinear with respect to x, but is still linear with respect to its parameters $\boldsymbol{\beta}$. Therefore quadratic, cubic and higher-order regressions (and indeed other basis functions that can be used with (10)) are often referred to as linear regression, depending on the context.

Penalised Splines

One approach to smoothing nonlinear data is the *penalised spline*. Penalised splines are also referred to as *semiparametric regression*, given that they are an extension of linear parametric regression (linear in the parameters), but also have the capabilities of nonparametric regression (i.e. local polynomial regression), such as the flexibility to accommodate nonlinear trends in the data (Ruppert et al, 2003).

The basis function which is the heart of the spline model is given by $(x - \kappa_h)_+^p$, where the "+" subscript denotes the positive part; in other words, for any number u, $u_+ = u$ if u is positive, and equals zero otherwise. The κ_h parameter is called the "knot" of the basis function; it is a value of x at which the spline is "split". Splines are constructed by using a number $h = 0, 1, 2, ..., H$ of spline basis functions with different knots κ_h. Specifically, the polynomial $\beta_0 + \beta_1 x_j + \cdots + \beta_p x_j^p$ in (9) is extended to give

$$\beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk} (x - \kappa_k)_+^p \tag{12}$$

where again $\beta.$ are coefficients to be estimated.

Since the model is linear with respect to its coefficients, it is possible to write it in the same form as (9) replacing $\mathbf{x}_j = (1, x_j, x_j^2, ..., x_j^p)^\top$ with $\mathbf{x}_j = (1, x_j, x_j^2, ..., x_j^p, (x_j - \kappa_1)_+^p, \ldots, (x_j - \kappa_K)_+^p)^\top$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{p1}, \ldots, \beta_{pK})^\top$. When $p = 1$, this is known as a *linear spline*. In practice, quadratic or cubic splines are normally used (corresponding to $p = 2$ or 3 respectively), because they result in a smooth fit to the data. In the applications here a value of $p = 3$ is used.

The coefficients may be directly estimated by using (11), however this tends to result in a fit which is too "rough"—in other words it fluctuates too much and is drawn too much to individual data points rather than following the smoother underlying trend. This is known as overfitting. To overcome this problem, the estimator (11) can be constrained to limit the influence of the spline basis terms, resulting in a smoother fit.

This results in what is known as *penalised splines*. To do this, $\boldsymbol{\beta}$ is chosen to minimise the sum of squared residuals, under the constraint that $\sum_{h=1}^{H} \beta_{pk}^2 = \boldsymbol{\beta}^\top \mathbf{D} \boldsymbol{\beta} < C$, where C is a positive constant and $\mathbf{D} = \operatorname{diag}\left(\mathbf{0}_{(p+1)\times(p+1)}, \mathbf{I}_H\right)$. The constraint reduces the influence of the last additional H terms $\sum_{h=1}^{H} \beta_{ph}(x-\kappa_h)_+^p$ in (12) to avoid over-fitting the data, and results in the *penalised spline* model. The constraining of regression parameters shares similarities with other approaches such as *ridge regression* (see e.g. Hastie et al (2001)), and *LASSO* (Tibshirani, 1996).

The constrained minimisation problem is solved by

$$\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{D})^{-1} \mathbf{X}^\top \mathbf{y} \tag{13}$$

where λ plays the role of a Lagrange multiplier for the penalisation. High λ values result in a very smooth fit, whereas low values give a rough fit which is closer to interpolating the data points. The smoothing parameter λ needs to be optimised, and this is done using a cross-validation approach, in which a search algorithm is used to find the value of λ that minimises the average squared error resulting from removing a point from the set of training data and predicting at that point. The fact that penalised splines are only a short step from linear regression means that they can exploit well-known properties to give fast order-n algorithms for the calculation of cross-validation measures, therefore the fitting of a penalised spline can be very fast. More details on this can be found in Ruppert et al (2003) Section 5.3.

The spline model requires the specification of the knots κ_h; here they are placed at equally-spaced quantiles of the unique x_i values, and the number of knots is chosen from a set of candidate H values, as the number of knots which results in the best fit according to a measure of cross-validation. This procedure is described in more detail in Ruppert et al (2003) Section 5.5.

In the applications in this chapter, the spline models are deliberately constrained to be quite simple fits to the data. To do this, the spline is allowed to have $K = 0, 1, 2, 3, 4$ knots, where the optimal number is chosen as that which minimises the cross-validation measure. In the case of $H = 0$ the spline is reduced to the polynomial model, which in this work is cubic, since p is set to 3.

Finally, note that in the following work "penalised splines" will sometimes be referred to simply as "splines". All splines used in this investigation are penalised cubic splines.

Local polynomial regression

Another approach to nonlinear regression is *local polynomial regression*, also referred to as *kernel smoothing*—see e.g. Bowman and Azzalini (1997). Kernel smoothing works by averaging a number of weighted polynomial regressions, centred at different values of x. The regression $\widehat{f}(x)$ is chosen here to be *local linear*, and the estimation is performed minimizing the weighted sum of squares, where weights are proportional to a kernel function with bandwith b. The kernel function gives the strongest weight to squared residuals corresponding to points close to x_j for observation j, which reflects the belief that the closer two points are to each other in x, the more likely they are to have similar values in y.

A commonly-used kernel function is the Gaussian density function with standard deviation b. Local linear regression is used here (as opposed to e.g. local mean regression) since it is generally regarded as a good choice, due to its good properties near the edges of the data cloud. The smoothing parameter b can be optimised by cross-validation; this is the method that is adopted in the applications in this chapter. The local polynomial approach is used by Da Veiga et al (2009) for estimation of first-order sensitivity indices of uncertainty for models with correlated inputs.

Remarks

The choice of whether to use penalised splines, kernel regression, linear regression, or another nonlinear regression approach entirely, is a working decision of the analyst. By construction, linear regression (i.e. linear with respect to x) only reveals linear dependencies between variables, so should be used with this caveat in mind. Higher-order polynomial regression or other basis functions within a linear regression might improve the fit, but also risks overfitting. The penalised splines and local-linear regression allow for nonlinearities when they are present in the data, but can also model near-linear data when required. In most applications, these two latter approaches are expected to give comparable results. However, in the presence of highly skewed and/or heteroskedastic data, the two fits may be quite different.

While the fits are likely to be similar, splines may have a slight advantage over local-linear regression in some other respects. First, they are less computationally demanding than kernel regression, due to the fast computation of cross-validation measures discussed previously. While the difference is small enough to be negligible for running a few regressions on small datasets, the difference may be sizable if one wants to attempt an optimisation of weights or if a very large number of regressions needs to be run. Large data sets of the order of thousands or millions of points are common in environmental science. A second advantage of splines is that the computation of their derivatives is just as easy.

In any case, a prudent strategy would be to run both analyses with splines *and* kernel regression, and compare the results — this is the approach taken in the following examples where both methods are applied and compared.

Case Studies

This section delves into the conceptual and statistical properties of three composite indicators selected for their interesting structure as well as for their popularity. These examples allow a practical demonstration of the methods described in previous sections, as well as providing an interesting analysis of three well-known composite indicators. The case studies are chosen because:

- they are transparent—namely the values of each input variable are publicly available, and the index can be reproduced given the methodological notes provided by the developers;
- they use three different aggregation formulas, which allows the effect of the different measures of importance described here, and estimation methods, to be assessed;
- they deal with issues of governance and accountability, and hence offer interesting narratives on the misconception of what matters when developing an index.

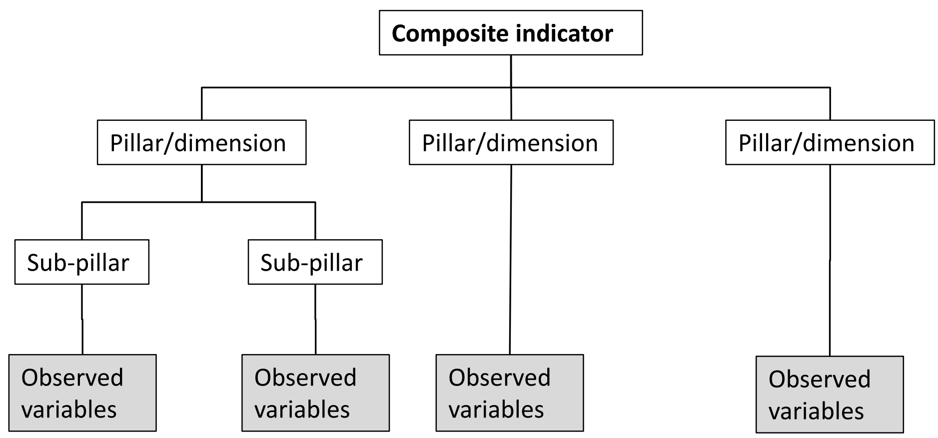


Fig. 2: A (fictional) example of the hierarchy of a composite indicator.

Often, composite indicators are built on a number of hierarchical "levels". Rather than having d measured variables $x_1, x_2, ..., x_d$ as direct inputs to the composite indicator, variables (indicators) are usually put together into groups, known as "pillars" or "dimensions", which share similar conceptual characteristics (see Figure 2). Variables in each pillar are aggregated in a weighted sum, such that each pillar is itself a composite indictor characterising one aspect of the greater theme. The composite values of each pillar are then used as the inputs of the composite indicator itself. Indeed, pillars may also consist of sub-pillars if the developers deem it appropriate. For simplicity however, and to be consistent with the sensitivity analysis literature and previous sections, the direct inputs to the composite indicators here are referred to as "variables" in the analytical parts of the following section, even though they might represent pillars.

Resource Governance Index (RGI)

Aim

The Resource Governance Index (RGI) is developed by the Revenue Watch Institute in order to measure the transparency and accountability in the oil, gas and mining sectors in 58 countries (Quiroz and Lintzer, 2013). These nations produce 85 percent of the world's petroleum, 90 percent of diamonds and 80 percent of copper, generating trillions of dollars in annual profits. The future of these countries depends on how well they manage their natural resources.

Sources

To evaluate the quality of governance in the extractive sector, the Resource Governance Index employs a 173-item questionnaire that is based on the standards put forward by the International Monetary Fund's 2007 *Guide on Resource Revenue Transparency* and the Extractive Industries Transparency Initiative, among others. The answers to the 173 questions are grouped into 45 indicators that are then mapped into three (of the four) RGI dimensions: Institutional and Legal Setting, Reporting Practices, and Safeguards and Quality Controls. The fourth dimension, Enabling Environment, consists of five additional indicators that describe a country's broader governance environment; it uses data compiled from over 30 external sources by the Economist Intelligence Unit, International Budget Partnership, Transparency International and Worldwide Governance Indicators. The Index is therefore a hybrid, with three dimensions based on the questionnaire specifically assessing the extractive sector, and the fourth rating the country's overall governance.

Main Dimensions

The RGI's four dimensions cover the following topics:

1. Institutional & Legal Setting (20% weight): 10 indicators that assess whether the laws, regulations and institutional practices enable comprehensive disclosures, open and fair competition, and accountability;
2. Reporting Practices (40% weight): 20 indicators that evaluate the actual disclosure of information and reporting practices by government agencies;
3. Safeguards and Quality Controls (20% weight): 15 indicators that measure the checks and oversight mechanisms that guard against conflicts of interest and undue discretion, such as audits; and
4. Enabling Environment (20% weight): 5 indicators of the broader governance environment generated using over 30 external measures of accountability, government effectiveness, rule of law, corruption and democracy.

The RGI score is a weighted arithmetic average of the four dimensions, i.e. of the form of (1), where the dimensions here are treated as the x_1, x_2, x_3, x_4 input variables.

Because actual disclosure constitutes the core of transparency, developers assigned a greater weight to the Reporting Practices dimension. This choice also reflects a belief that without reporting information, rules and safeguards ring hollow. Therefore, Reporting Practices are assigned a weight equal to 40% of the final country score, and the other three dimensions (Institutional and Legal Setting, Safeguards and Quality Controls and Enabling Environment) account for 20 percent each.

On the inclusion of the Enabling Environment dimension, there have been arguments for and against. Against its inclusion, the arguments are:

1. the Enabling Environment dimension dilutes the focus of the index on the oil, gas and mining sector by incorporating measures of overall governance;
2. The Enabling Environment dimension can have an undue effect on the index scores, driving scores up or down, inflating or depressing performances beyond what countries actually show in their extractive sector.

In favour of its inclusion, the arguments are:

1. External governance indicators reflect the influence of the broader country environment on the quality of natural resource governance. When considering the quality of transparency and accountability in a certain area, it does matter whether a country also has an authoritarian regime, a high risk of corruption or respect for basic freedoms.
2. As an expert-based index, the accuracy and consistency of its findings suffer from the bias introduced by researchers, and by peer and Research Watch Institute reviewers. Including this dimension as an external measure reduces this margin of error.

Given these last two arguments, the developers chose to include the Enabling Environment dimension and to allocate a 20 percent weight to this dimension. As part of the Index website, the developers provide a tool that allows users to change the weights for the different dimensions, creating different composite scores that reflect their own sense of prioritisation. This is a direct example of how weights are often (erroneously) interpreted as measures of importance.

Results

Before examining the relationships between each variable (dimension/pillar) and the output of the composite indicator, a basic analysis of the structure of the data was performed. There are no outliers (absolute skewness is less than 0.63) in the four dimensions' distributions that could bias the subsequent interpretations of the correlation structure. The four dimensions of the Resource Governance Index have moderate to high correlations that range from 0.41 (between Institutional and Legal Setting and Enabling Environment) to 0.82 (between Reporting Practices and Safeguards and Quality Controls) and an overall good average bivariate correlation of 0.65. Principal component analysis suggests that there is indeed a single latent phenomenon underlying the four index dimensions. This first principal component captures 74 percent of the variation in the four dimensions.

Moving now to the importance measures, Table 1 shows the estimates of the correlation ratios obtained both with penalised splines and local-linear (LL) regression, as well linear correlations, for the four input variables of the index. The correlation ratios S_i (penalised splines and LL approach) and the Pearson correlation coefficients R_i^2 confirm that the Reporting Practices component has indeed the highest impact on the index. This was the intention of the RGI developers on the grounds that actual disclosure constitutes the core of transparency. This choice also reflects a belief that without reporting information, rules and safeguards are inconsequential. In fact, the correlation between Reporting

Resource Governance Index ($n = 58$)	x_i	w_i	R_i	R_i^2	$S_{i,\mathrm{spl}}$	$S_{i,\mathrm{LL}}$
Institutional and legal setting	x_1	0.2	0.79	0.63	0.65	0.67
Reporting practices	x_2	0.4	0.95	0.90	0.90	0.94
Safeguards and quality controls	x_3	0.2	0.91	0.82	0.83	0.83
Enabling Environment	x_4	0.2	0.77	0.59	0.65	0.70

Table 1: Measures of dependence of the Resource Governance Index on its input variables: $R_i = \mathrm{corr}(x_{i,}, y)$: correlation; $S_{i,\mathrm{spl}}$: correlation ratio, penalised spline; $S_{i,\mathrm{LL}}$: correlation ratio, local-linear.

Practices and Safeguards and Quality Control is very high (0.82, the highest among the components). If one could fix the Reporting Practices variable, the variance of the RGI scores across the 58 countries would on average be reduced by 94% (local-linear estimate). It is worth noting that despite the equal weights assigned to the other three components, their impact on the RGI variation differs: by fixing any of the other variables, the variance reduction would be 83% for Safeguards and Quality Control, 70% for Enabling Environment and 67% for Institutional and Legal Setting, using the estimates of the local-linear regression.

As per the developers intention to have Reporting practices twice as important as any of the other three dimensions, i.e. Institutional and legal setting, Safeguards and quality controls, and Enabling Environment, it is easy to see that this was not achieved. The strong correlation of Reporting practices with Safeguards and quality controls results in these two pillars being almost equally important (0.90 and 0.82), while the importance of Reporting practices relative to the remaining two pillars is of about 3/2 rather than 2.

Aside from the implications of this analysis from the point of view of the RGI, it is interesting to look at the differences between the measures of importance in Table 1. The linear R_i^2 measure consistently gives the lowest estimates of importance, the LL estimate of correlation ratio the highest, and the penalised spline estimate of correlation ratio is somewhere in between. To see in a little more detail what is happening, Figure 3 shows the scatterplots of the four input variables and the simple linear, penalised spline and local-linear estimations of main effects. For the Institutional and Legal Setting variable, the two nonlinear fits are significantly different to the linear fit, but fairly similar to each other. However the structure of the data is such that the gradient of the linear fit (and the

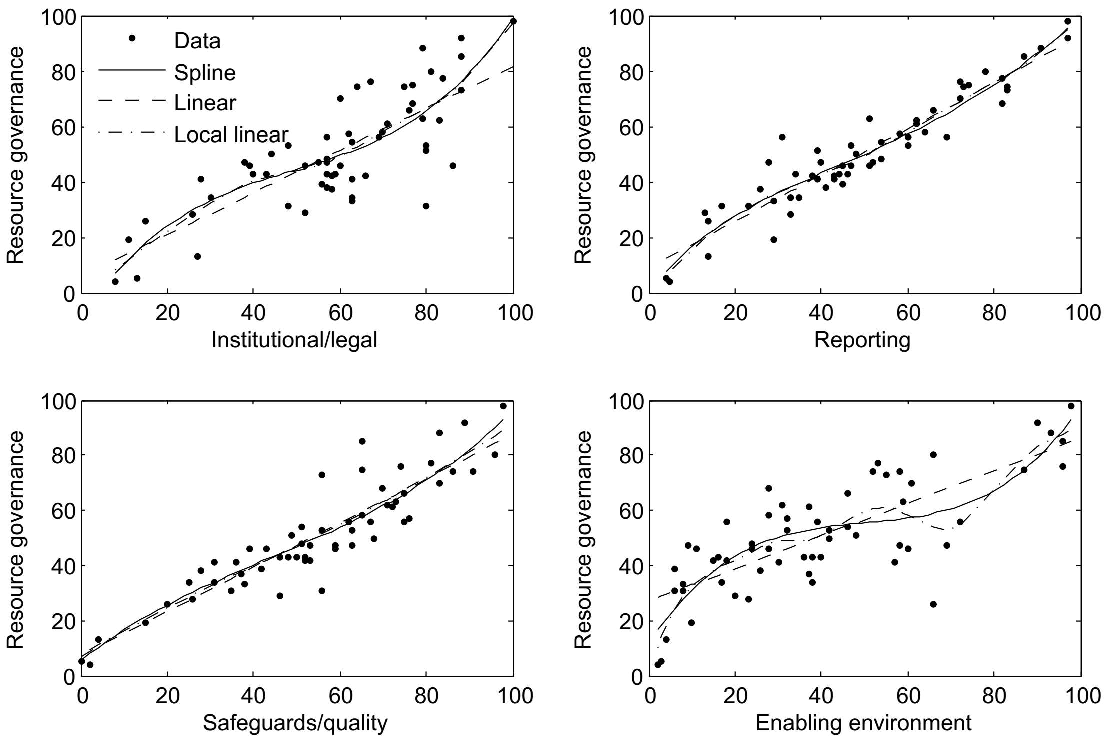


Fig. 3: Penalised spline, local-linear and linear fits to the Resource Governance Index.

overall trend of the nonlinear fits) is not very steep, resulting in relatively low importance estimates compared to other variables. The Reporting Practices variable shows quite similar fits between all three methods, although the local-linear curve has a slightly higher variance. Safeguards and Quality gives a near-linear fit for both nonlinear regression approaches, showing a strong agreement in measures of importance. The Enabling Environment variable is the most interesting here from a regression point of view: the penalised spline fits a roughly cubic model, whereas the local-linear curve fluctuates more strongly with the data. Whether the more parsimonious penalised spline curve, or the more variable local-linear curve is a better estimate of $E(y|x_i)$ is not intuitively clear from visual inspection.

As a further analysis of the effect of each variable, Figure 4 shows the first derivatives of $E(y|x_i)$ with respect to each input variable, as estimated by the penalised splines. First, the nonlinearities of the spline fits are evident from the non-constant derivatives. The derivative plots also have implications for the index itself—they effectively summarise the expected change in the RGI that a country would achieve if it moved a given amount in each variable. Two example countries, Libya and India, have their scores in each indicator marked as dotted vertical lines. In the case of Libya, it is clear that to move up the rankings in the RGI, it would better invest efforts in improving the Institutional and

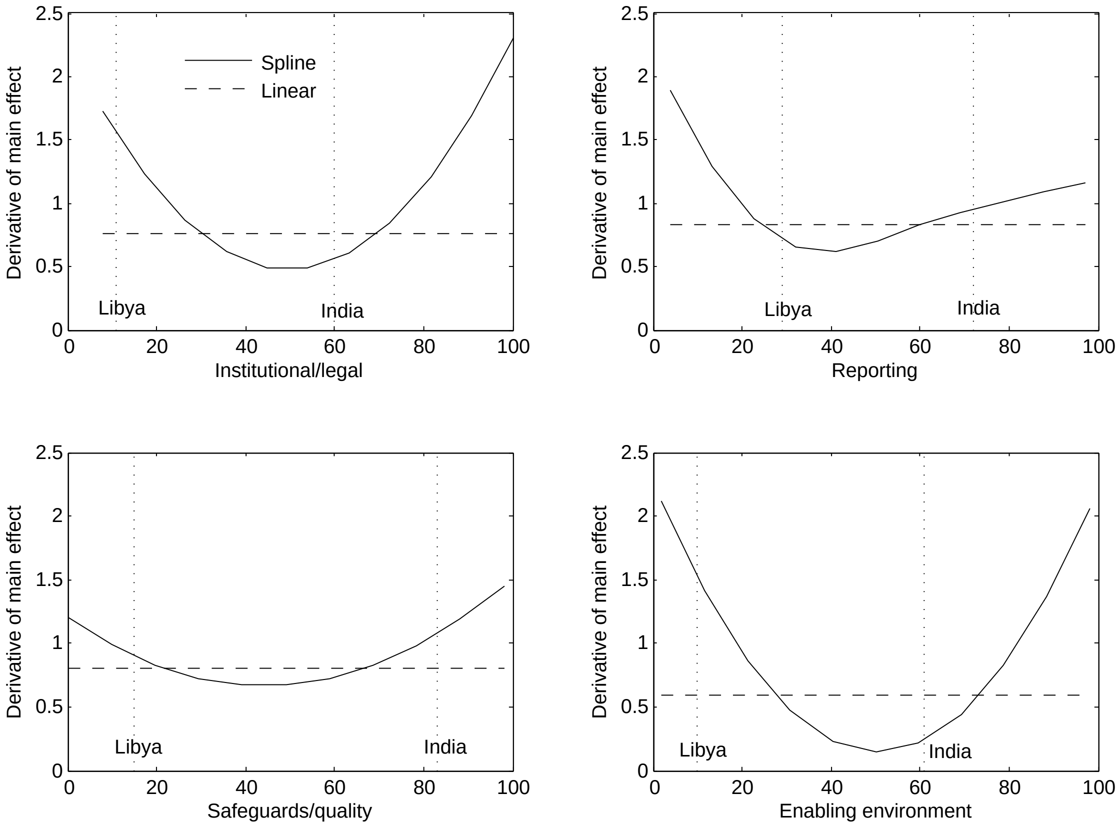


Fig. 4: Derivatives $d[E(y|x_i)]/dx_i$ using penalised splines and linear regression fits to the Resource Governance Index. Indicator values of Libya and India marked as vertical dotted lines.

Legal Setting, and Enabling Environment dimensions, whereas gains in Reporting and Safeguards and Quality would yield lesser gains. India, a country ranked overall 12th in the RGI, would on the other hand stand to gain very little from small improvements in Enabling Environment, and immediate efforts would be better directed towards either Reporting or Safeguards and Quality.

Financial Secrecy Index (FSI)

Aim

The Financial Secrecy Index (FSI) is developed by the Tax Justice Network (TJN) and aims to measure the contribution to the global problem of financial secrecy in 80 jurisdictions worldwide (Cobham et al, 2013). In other words, the Index attempts to provide an answer to the question: by providing offshore financial services in combination with a lack of transparency, how much damage is each secrecy jurisdiction actually responsible for?

To give an example of what this implies, the home page of the FSI informs us that because of capital flights dwarfing the inflow of aid money Africa is in fact a net creditor to the rest of the world since the seventies. As per the money stashed away the TJN informs us that "those assets are in the hands of a few wealthy people, protected by offshore secrecy, while the debts are shouldered by broad African populations".

Sources

The index combines qualitative data and quantitative data. Qualitative data are based on laws, regulations, cooperation with information exchange processes and other verifiable data sources and are used to calculate a secrecy score for each jurisdiction. Secrecy jurisdictions with the highest secrecy scores are more opaque in the operations they host, less engaged in information sharing with other national authorities and less compliant with international norms relating to combating money-laundering. Lack of transparency and unwillingness to engage in effective information exchange makes a secrecy jurisdiction a more attractive location for routing illicit financial flows and for concealing criminal and corrupt activities.

Quantitative data is used to create a global scale score, for each jurisdiction, according to its share of offshore financial services activity in the global total. Publicly available data about the trade in international financial services of each jurisdiction are used. Where necessary because of missing data, the developers follow International Monetary Fund methodology to extrapolate from stock measures to generate flow estimates. Jurisdictions with the largest weightings are those that play the biggest role in the market for financial services offered to non-residents.

Main Dimensions

The first dimension of the FSI is the Financial Secrecy score, which is calculated from a set of fifteen qualitative key financial secrecy indicators that assess the degree to which the legal and regulatory systems (or their absence) of a country contribute to the secrecy that enables illicit financial flows. These indicators can be grouped around four dimensions of secrecy: 1) knowledge of beneficial ownership (3 indicators); 2) corporate transparency (3 indicators); 3) efficiency of tax and financial regulation

(4 indicators); and 4) international standards and cooperation (5 indicators). Taken together, these indicators result in one aggregate secrecy score for each jurisdiction.

The second dimension of the FSI is the Global Scale score, which is calculated based on quantitative data (publicly available) about the trade in international financial services, and captures the potential for a jurisdiction to contribute to the global problem of financial secrecy. Data on international trade in financial services come from the IMF's Balance of Payments statistics. Data on stocks of portfolio assets and liabilities are taken from two IMF sources: the Coordinated Portfolio Investment Survey and the International Investment Position statistics.

At the final step, the Financial Secrecy score is cubed and the Global Scale is cube-rooted before being multiplied to produce the FSI scores for each jth jurisdiction, i.e.

$$FSI_j = \text{Secrecy}_j^3 \cdot \sqrt[3]{\text{GlobalScale}_j}.$$

Critics have argued that the Global Scale dimension unfairly points to large financial centres. However, the developers response is that "to dispense with the scale risks ignoring the big elephants in the room". While large players may be slightly less secretive than other jurisdictions, the extraordinary size of their financial sectors offers far more opportunities for illicit financial flows to hide. Therefore, the larger an international financial sector becomes, the better its regulations and transparency ought to be. This logic is reflected in the FSI which aims to avoid the conceptual pitfalls of the "usual suspects"—lists of tax havens which are often remote islands whose overall share in global financial markets is tiny. In the FSI a jurisdiction with a larger share of the offshore finance market, and a moderate degree of opacity, may receive the same overall ranking as a smaller but more secretive jurisdiction.

Due to a significantly greater skew in the Global Scale scores compared to the Financial Secrecy scores, the developers transform the two to generate series that are more evenly distributed. The choice of the transformation has been guided by the 90/10 and the 75/25 percentile ratios in each of the two series. In the original series, the 90/10 percentile ratio is more than five thousand times higher for the Global Scale than the Secrecy variable; the 75/25 ratio nearly a hundred times higher. If one squares the Secrecy Score and takes the square root of the Global Scale, these ratios fall to below 26 and 6 respectively; and if one cubes the Secrecy Score and takes the cube root of the Global Scale, they fall below 3 and 2 respectively. Finally, looking at fourth and fifth roots and powers, the variation of the Global Scale series become disproportionately small. Hence, the cube root/cube

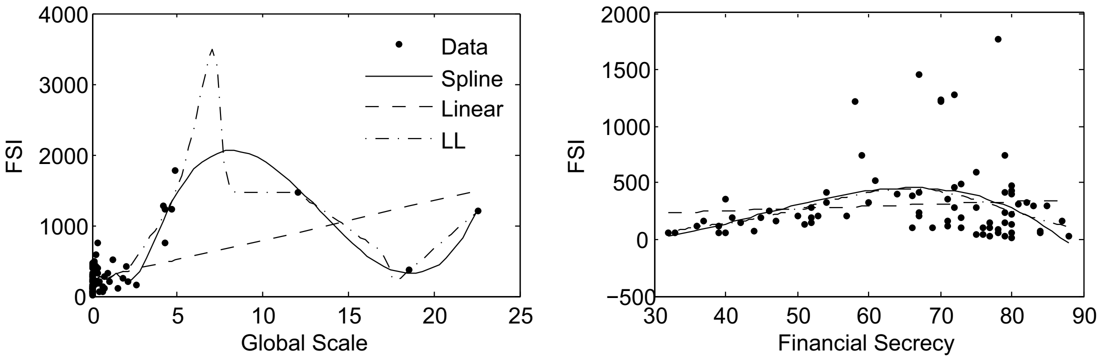


Fig. 5: Scatter plots of Financial Secrecy Index versus its components.

Financial Secrecy Index (n=80)	x_i	w_i	R_i	R_i^2	$S_{i,\mathrm{spl}}$	$S_{i,\mathrm{LL}}$
Global Scale	x_1	0.5	0.57	0.32	0.63	0.79
Financial Secrecy	x_2	0.5	0.09	0.01	0.13	0.09

Table 2: Importance measures of the variables of the Financial Secrecy Index. $R_i = \mathrm{corr}(x_{i,}, y)$: correlation; $S_{i,\mathrm{spl}}$: correlation ratio, penalised spline; $S_{i,\mathrm{LL}}$: correlation ratio, local-linear.

combination was adopted by the developers on the grounds that "these transformations are sufficient to ensure that neither secrecy nor scale alone determine the FSI"—see Cobham et al (2015).

Results

Despite the intentions of the developers, and the reasoning based on the quantiles, the correlation ratios S_i (splines and LL approach) and the Pearson correlation coefficients reveal a notably unbalanced impact of the two components on the FSI (see Table 2). The greatest difference between the estimates is in the correlation ratios provided by the local-linear regression, in which the values are 0.79 for Global Scale and 0.09 for secrecy. This is in stark contrast to the intended influence, which should be roughly equal for both variables. Examining the scatterplots in Figure 5 however, the data looks quite challenging to smooth, and the three regression approaches have markedly different fits. In particular, the local-linear regression of the Global Scale variable has a large spike in the fit at a value of around 7, which does not appear to be justified by the data. Therefore the LL importance estimates ought to be treated with caution. The spline fit is slightly more convincing, but seems to be quite heavily biased by the outlying points above a score of around 5. However given that even the linear fit yields a far higher importance measure for the Global Scale variable than the Secrecy variable (0.32 and 0.01 respectively), the evidence seems fairly compelling that the Global scale variable dominates the FSI by a significant margin. The analysis illustrates vividly that the cube root/cube transformation of the FSI components and the equal weights assigned to the two components are not a sufficient condition to ensure equal importance, at least according to the correlation ratio measure.

The Global scale scores are particularly skewed (skewness = 4.6, and kurtosis = 23). Yet, after considering the cube-root, the distribution is no longer problematic (skewness = 1.7, kurtosis = 3.0). Nevertheless, what remains problematic is the strong negative association between the cubed distribution of the Financial Secrecy scores and the cube-rooted distribution of the Global Scale scores (corr = $-.536$).

If the points with the highest Global Scale scores are treated as outliers, the problem lessens somewhat but does not disappear. Table 3 shows the results of re-running the analysis, trimming the points with three highest or eight highest Global Scale values. By visual inspection, these seem to be two possible reasonable definitions of outliers depending on where the cutoff value is set. Trimming the top three points, the disparity is very similar, with the Global Scale correlation ratios far higher than the Secrecy values. After trimming the top eight points, the importance scores are much closer, however in the linear and spline estimates, the importance of the Global Scale variable is still at least twice that of the Secrecy variable. In the case of the LL regression, the Secrecy variable is now rated as slightly more influential than the Global Scale, but given that this is the only (mild) contradiction to an otherwise compelling trend, the conclusion that the variables are not equally influential appears to stand.

In defense of the developers one may note that this gross unbalance refers to a specific definition of importance (how much the variance of the index would be reduced on average if one could fix one dimension), and this definition appears to condemn the FSI as problematic. One might argue however that the use of variance of the main effect $E(y|x_i)$ as a measure of importance might not tell the whole story, and that interactions between the two variables should also be accounted for. Following this line of thought, a useful tool might be the "total effect index" S_{Ti}, which is defined as $E_{\boldsymbol{x}_{\sim i}}[V_{x_i}(y \mid \boldsymbol{x}_{\sim i})]/V(y)$, where $\boldsymbol{x}_{\sim i}$ denotes the vector of all variables but x_i. This measure also accounts for variance due to interactions. In fact the scatterplot-plot of Financial Secrecy in Figure 5

N_{trim}	Financial Secrecy Index (n=50)	R_i^2	$S_{i,\mathrm{spl}}$	$S_{i,\mathrm{LL}}$
3	Global Scale	0.616	0.798	0.754
	Financial Secrecy	0.017	0.112	0.074
8	Global Scale	0.028	0.310	0.077
	Financial Secrecy	0.015	0.119	0.093

Table 3: Importance measures of the variables of the Financial Secrecy Index with the N_{trim} points with the highest Global Scale scores removed. $R_i = \mathrm{corr}(x_{i,}, y)$: correlation; $S_{i,\mathrm{spl}}$: correlation ratio, spline; $S_{i,\mathrm{LL}}$: correlation ratio, local-linear.

looks like a textbook example of a variable with a low S_i and a high S_{Ti}, with the points displaying a rather constant mean and an increasing variance while moving over the horizontal axis.

An investigation using this measure in the context of composite indicators is outside of the scope of this work, but further information can be found in the section on Variance-based Sensitivity Analysis: Theory and Applications.

Good Country Index

Aim

The Good Country Index (GCI) is developed by the Good Country Party with a view to measure what a country contributes to the common good of humanity, and what it takes away (Anholt and Govers, 2014), following its developers' normative framework and world view. In total, 125 countries are included in the Index. In contrast to the majority of similar composite indicators, the Good Country Index does not measure what countries do at home; rather, it aims to start a global discussion about how countries can balance their duty to their own citizens with their responsibility to the wider world.

This reflects the consideration that the biggest challenges facing humanity today are global and borderless: problems such as climate change, economic crisis, terrorism, drug trafficking, slavery, pandemics, poverty and inequality, population growth, food and water shortages, energy, species loss, human rights and migration.

All of these problems stretch across national borders, so they can be properly tackled through international efforts. Hence, the concept of the "Good Country" is about encouraging populations and their governments to be more outward-looking, and to consider the international consequences of their national behaviour.

Sources

The GCI builds upon 35 indicators that are produced by the United Nations and other international agencies, and a few by NGOs and other organisations. Most of the indicators used are direct measurements of world-friendly or world-unfriendly behaviour (such as signing of international treaties, pollution, acts of terrorism, wars) and some are rather indirect (such as Nobel prizes, exports of scientific journals). By adding them up, the developers aim to get a reasonable picture of whether each country is effectively a net creditor to the rest of humanity in each of the seven categories, or whether it is a “freeloader” on the global system and ought to be recognised as such.

Main Dimensions

The 35 indicators are split in seven groups of five indicators each. These seven dimensions, which closely mirror the dimensions of the United Nations Charter, are:

1. Science, Technology & Knowledge, which includes foreign students studying in the country; exports of periodicals, scientific journals and newspapers; articles published in international journals; Nobel prize winners; and International Patent Cooperation Treaty applications.
2. Culture, which measures exports and imports of creative goods; UNESCO dues in arrears (a negative indicator); countries and territories that citizens can enter without a visa; and freedom of the press (based on the mean score of the Reporters without Borders and Freedom House indices as a negative indicator).
3. International Peace and Security, which aggregates peacekeeping troops sent overseas; dues in arrears to financial contribution to UN peacekeeping missions (negative indicator); casualties of international organised violence (negative indicator); exports of weapons and ammunition (negative indicator); and the Global Cyber Security Index (negative indicator).
4. World Order, which measures population that gives to charity as proxy for cosmopolitan attitude; refugees hosted; refugees overseas (negative indicator); population growth rate (negative indicator); and treaties signed as proxy for diplomatic action and peaceful conflict resolution.
5. Planet and Climate, which measures the National Footprint Accounts Biocapacity reserve; exports of hazardous waste (negative indicator); organic water pollutant emissions (negative indicator);

CO_2 emissions (negative indicator); and methane + nitrous oxide + other greenhouse gases (HFC, PFC and SF6) emissions (negative indicator).

6. Prosperity and Equality, which aggregates trading across borders; aid workers and volunteers sent overseas; fair trade market size; Foreign Direct Investment outflow; and development cooperation contributions (aid).
7. Health and Wellbeing, which includes wheat-tonnes-equivalent food aid shipments; exports of pharmaceuticals; voluntary excess contributions to the World Health Organisation; humanitarian aid contributions; and drug seizures.

A ranking is calculated for each of the seven dimensions. The Good Country Index is then calculated by taking the arithmetic average of the seven ranks in Science, Technology and Knowledge; Culture; International Peace and Security; World Order; Planet and Climate; Prosperity and Equality; and finally Health and Wellbeing. This aggregation scheme has been selected by the developers because of its simplicity and relative robustness to outliers. Beyond what is stated by the developers, a further argument in favour of this aggregation scheme would be the "imperfect substitutability across all seven index components", i.e. the reduction of the fully compensatory nature of an arithmetic average of the seven scores.

Results

Before looking at the correlation ratios describing the importance of each variable to the output, it is useful to look at the correlations between input variables, as it is good practice in the construction/evaluation of composite indicators. Six of the seven dimensions of the Good Country Index have low to moderate correlations that range from 0.20 (between several pairs of dimensions, mostly those involving Prosperity and Equality) to 0.78 (between Science and Technology and Culture) and an overall moderate average bivariate correlation of 0.37. Principal component analysis suggests that there is indeed a single latent phenomenon underlying the six dimensions, and that the first principal component captures almost 50% of the variation in these dimensions. However, the International Peace and Security dimension has a negative correlation to both the Science and Technology and to the Culture variables (-0.48), and almost random correlation to all remaining dimensions. This point is a strong concern for the validity of the GCI. The negative correlations between International Peace and Security on one hand and either Science and Technology, or Culture or Health and Wellbeing on

Good Country Index (n=125)	x_i	w_i	R_i	R_i^2	$S_{i,\text{spl}}$	$S_{i,\text{LL}}$
Science and Technology	x_1	1/7	0.71	0.50	0.50	0.50
Culture	x_2	1/7	0.79	0.63	0.63	0.66
International Peace and Security	x_3	1/7	−0.17	0.03	0.05	0.03
World Order	x_4	1/7	0.78	0.62	0.64	0.63
Planet and Climate	x_5	1/7	0.57	0.32	0.34	0.33
Prosperity and Equality	x_6	1/7	0.49	0.24	0.27	0.25
Health and Wellbeing	x_7	1/7	0.55	0.30	0.35	0.37

Table 4: $R_i = \text{corr}(x_{i,}, y)$: correlation; $S_{i,\text{spl}}$: correlation ratio, spline; $S_{i,\text{ker}}$: correlation ratio, local-linear.

the other, are undesirable in a composite indicator context, as they suggest the presence of trade-offs, and are a reminder of the danger of compensability between components.

Turning now to the correlation coefficients, Table 4 shows that, unlike the equal weighting scheme of the seven components would suggest, the impact of the seven components on the index is uneven. Three of the seven components, namely Culture, World Order, and Science and Technology, account for over 50% of the variation in the index scores (up to 63% for Culture). Instead, by fixing either of the three components—Health and Wellbeing, Planet and Climate, and Prosperity and Equality—the variance reduction would be between 25-37%. What is striking is that the International Peace and Security component is practically cosmetic in the framework: by fixing this component, the index variance would be reduced by merely 3%. Moreover, it actually has a negative correlation with the GCI output, meaning that countries that rank low in International Peace and Security, on average, actually stand out as "good countries", with a higher GCI score. This effect, likely due to the negative correlations with other variables, suggests a weakness in the GCI which ought to be addressed.

This conclusion can be of value in a general sense because it indicates that, aside from the choice of weights and aggregation formulas based on subjective considerations, the impact of components on the variance of the index is driven by the statistical properties of the components, and this latter fact is often overlooked during the index development process.

Discussion on estimation approaches

The tables showing correlation ratios and correlation coefficients from the case studies in the previous section (i.e. Tables 1, 2 and 4) help to understand the relative merits of the measures used in this study. The correlation coefficient R_i and the coefficient of determination R_i^2 give the linear correlation of the composite indicator with each of its inputs. The R_i^2 measure can be interpreted as the fraction of variance accounted for by the linear regression (similar to S_i), but it is also useful to see the R_i value in order to understand whether the relation is positive or negative. As shown in the Good Country Index (Table 4), this measure revealed that one of the input indicators is in fact negatively correlated with the composite indicator—this is an undesirable property as discussed previously, at least in the context of linearly or geometrically aggregated composite indicators.

The spline and local-linear estimates of S_i are all higher than or equal to the R_i^2 values. In most cases the difference is small — this reflects the fact that the main effects are close to linear. In these cases (see for example the "Reporting Practices" variable of the Resource Governance Index in Table 1), the spline and LL regression fits are very close to straight lines. However, in some cases, such as the Enabling Environment pillar (also from the Resource Governance Index), S_i estimates are significantly higher than R_i^2 — this is due to the nonlinear main effect, which is captured by the spline and LL regression, but missed by the linear regression. In this instance, $R^2 = 0.59$ and S_i is estimated as 0.65 or 0.70 by the spline and LL regression respectively. Looking at all three tables, one can see that in the majority of cases R_i^2 would be a sufficient approximation of the effect of each variable, but there are several exceptions. This shows the added value of nonlinear regression—it can approximate linear and near-linear main effects, which appear in the majority of cases, but can generalise to nonlinear fits when required.

An obvious question at this point is to ask whether splines or LL regression provides a better estimate of S_i, given that the fits are slightly different in some cases. The short answer is that there is no way to know which is better, given that we do not know the "true" values of S_i or the "true" main effects. The three tables show however that in the majority of cases, the estimates are very similar, and in the cases where the estimates differ, neither the splines or the LL regression has a tendency to provide higher estimates than the other overall.

It is tempting to think that higher estimates might be better estimates, given that they "capture" more of the total variance. However, both the splines and the LL regression could be easily adjusted to capture 100% of the variance; this would result in interpolation rather than smoothing. But this does not in general provide a good approximation of the main effects, and results in *overfitting*. Instead the aim of nonlinear regression is to find a balance between interpolation and simplicity — this is known as the "bias-variance tradeoff" in machine learning; see e.g. Hastie et al (2001).

Given then that neither splines nor local-linear regression provide consistently higher or lower estimates than the other, the best strategy would be to estimate main effects using both methods and then compare the results. It can also be helpful to visually examine the fits – for example in Figure 5, the Global Scale variable of the Financial Secrecy Index gives a dataset that results in very different main effect estimates between the linear, spline and local-linear approaches. Although none of the fits seem extremely convincing (the data is strongly skewed and heteroskedastic), the LL fit in particular has a strange "spike" that does not appear to be justified by the data—in this case one might treat the estimate of the local-linear regression with caution. A clear avenue of research here would be to incorporate confidence intervals into the estimates of main effects, and subsequently into the estimates of S_i. Some approaches to doing this within the context of splines can be found in Ruppert et al (2003). An alternative tool might be to use a Bayesian implementation of a Gaussian process, which would formally propagate uncertainties in main effect estimation to estimates of correlation ratios.

Aside from the fitting of the data, splines may offer some small advantages over LL regression. The first is that, being closely related to linear regression, they can provide fast analytic estimates of derivatives. This property is used in Figure 4, to illustrate the gain in Resource Governance Index that a country or entity would achieve if its value of a given variable changed by a given amount along its axis. Furthermore, splines are able to exploit properties of linear regression (such as calculation of the cross validation measures) to allow very fast fitting to a given data set. In the examples here, which feature relatively small data sets, computational time is not an issue. However some composite indicators based on physical maps measure concepts such as ecosystems services indices at resolutions as high as 1km, over the whole of Europe (Paracchini et al, 2014). In such cases, the data set may feature millions of points, and splines offer a significant advantage over the slower local-linear regression.

Conclusions

The leitmotif of the present chapter is to "mind the gap" between the two stories which can be told about the weights of composite indicators. On one hand weights appear to "belong" to developers, who are entitled to set them by analysis and negotiation based on their norms and values. On the other hand, most forms of aggregation—and notably all linear aggregations—are particularly inept at translating these weights into "effects". The proposal here is to look at the problem by choosing a statistically defensible definition of importance, one that derives from the theory of global sensitivity analysis, which in turn originates from the theory of experimental design and of analysis of variance (ANOVA-like decomposition). It is therefore possible to compare the importance of different variables against what their weights would purport them to be. The discussion of the Financial Secrecy Index makes it clear that all inferences made here are conditional on the definition of "importance". There is nothing surprising in this. The same holds in sensitivity analysis. In fact *'in sensitivity analysis [an] analyst [must] specify what is meant by "factors importance" for a particular application. Leaving instead the concept of importance undefined could result in several tests being thrown at the problem and in several rankings of factor importance being obtained, without a basis to decide which one to believe'* (Saltelli et al, 2012). The situation is similar in the analysis of the coherence of the weights with the behavior of the index.

Still, within the limits of this analysis it is perhaps possible to say something useful to the developers. As far as the Resource Governance Index is concerned it can be said that developers were successful in making Reporting practices the most important dimension. They were less successful in making the other three dimensions equally important (Table 1). Further, in the intention of the developers each of these three dimensions "Institutional and Legal Setting", "Safeguards and Quality Controls", and "Enabling Environment", were supposed to be half important as "Reporting practices". The correlation structure of the problem did not allow this to happen, at least with the selected weights.

For the Financial Secrecy Index, the measures of importance point to a problematic property of the index, whereby one dimension, "Global Scale", apparently far more important than the other, "Financial Secrecy" (with the exact ratio varying somewhat between various estimates), while the two are intended to be equally important. At the same time, the aggregation formula chosen generates an important interaction term in the index. This challenges the correlation ratio as an sufficient measure of importance for this particular application. More analysis is needed on this challenging test case.

In the case of the Good Country Index, it is clear that the ambition to capture several dimensions normatively labeled as equally important was not rewarded by the results. Not only are the dimensions unequal, but one important dimension, "International Peace and Security" has a very small influence on the index score, and in fact has a negative correlation with the output.

Note that the approach presented in this chapter can be used to actually improve the indices by tuning the weights as to obtain the desired importance (Paruolo et al, 2013). The approach has been implemented in two important indices: the INSEAD-WIPO Global Innovation Index (Saisana and Saltelli, 2014) and the Environmental Performance Index developed by Yale University (Athanasoglou et al, 2014).

References

Anholt S, Govers R (2014) The Good Country Index. Tech. rep., The Good Country Party, URL http://www.goodcountry.org/

Athanasoglou S, Weziak-Bialowolska D, Saisana M (2014) Environmental performance index 2014: Jrc analysis and recommendations. Tech. rep., European Commission, Joint Research Centre

Bandura R (2011) Composite indicators and rankings: Inventory 2011. Tech. rep., United Nations Development Programme Office of Development Studies

Bowman A, Azzalini A (1997) Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations, vol 18. Oxford University Press, USA

Cobham A, Jansky P, Christensen J, Eichenberger S (2013) Financial Secrecy Index 2013: Methodology. Tech. rep., The Tax Justice Network, URL http://www.financialsecrecyindex.com/

Cobham A, Jansk P, Meinzer M (2015) The financial secrecy index: Shedding new light on the geography of secrecy. Economic Geography 91(3):281–303

Da Veiga S, Wahl F, Gamboa F (2009) Local polynomial estimation for sensitivity analysis on models with correlated inputs. Technometrics 51(4):452–463

Decancq K, Lugo MA (2013) Weights in multidimensional indices of wellbeing: An overview. Econometric Reviews 32(1):7–34

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Data mining, inference and prediction. Springer

Kelley JG, Simmons BA (2015) Politics by number: indicators as social pressure in international relations. American Journal of Political Science 59(1):55–70

Li G, Rabitz H, Yelvington PE, Oluwole OO, Bacon F, Kolb CE, Schoendorf J (2010) Global sensitivity analysis for systems with independent and/or correlated inputs. The Journal of Physical Chemistry A 114(19):6022–6032

Paracchini ML, Zulian G, Kopperoinen L, Maes J, Schägner JP, Termansen M, Zandersen M, Perez-Soba M, Scholefield PA, Bidoglio G (2014) Mapping cultural ecosystem services: A framework to assess the potential for outdoor recreation across the eu. Ecological Indicators 45:371–385

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

Pearson K (1905) On the General Theory of Skew Correlation and Non-linear Regression, volume XIV of Mathematical Contributions to the Theory of Evolution, Drapers' Company Research Memoirs. Dulau & Co., London, Reprinted in: Early Statistical Papers, Cambridge University Press, Cambridge, UK, 1948

Quiroz JC, Lintzer M (2013) The 2013 resource governance index. Tech. rep., The Revenue Watch Institute

Rasmussen C, Williams C (2006) Gaussian Processes for Machine Learning. The MIT Press, Cambridge, Massachusetts

Ruppert D, Wand M, Carroll R (2003) Semiparametric regression, vol 12. Cambridge Univ Pr

Saisana M, Saltelli A (2014) Joint Research Centre statistical audit of the 2014 Global Innovation Index. Tech. rep., European Commission, Joint Research Centre

Saisana M, Saltelli A, Tarantola S (2005) Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. Journal of the Royal Statistical Society - A 168(2):307–323

Saisana M, d'Hombres B, Saltelli A (2011) Rickety numbers: Volatility of university rankings and policy implications. Research policy 40(1):165–177

Saltelli A, Tarantola S (2002) On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. Journal of American Statistical Association 97:702–709

Saltelli A, Tarantola S, Campolongo F (2000) Sensitivity analysis as an ingredient of modelling. Statistical Science 15(4):377–395

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) Global Sensitivity Analysis - The Primer. John Wiley & Sons, Ltd

Saltelli A, Ratto M, Tarantola S, Campolongo F (2012) Update 1 of: Sensitivity analysis for chemical models. Chemical Reviews 112(5):1–25

Storlie C, Helton J (2008) Multiple predictor smoothing methods for sensitivity analysis: Description of techniques. Reliability Engineering & System Safety 93(1):28–54

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) pp 267–288

Yang L (2014) An inventory of composite measures of human progress. Tech. rep., United Nations Development Programme Human Development Report Office