Journal of the Royal Statistical Society
SERIES A Statistics in Society

J. R. Statist. Soc. A (2013)
176, *Part* 3, *pp.* 609–634

Ratings and rankings: voodoo or science?

Paolo Paruolo

University of Insubria, Varese, Italy

and Michaela Saisana and Andrea Saltelli

European Commission, Ispra, Italy

[Received March 2011. Final revision May 2012]

Summary. Composite indicators aggregate a set of variables by using weights which are understood to reflect the variables' importance in the index. We propose to measure the importance of a given variable within existing composite indicators via Karl Pearson's 'correlation ratio'; we call this measure the 'main effect'. Because socio-economic variables are heteroscedastic and correlated, relative nominal weights are hardly ever found to match relative main effects; we propose to summarize their discrepancy with a divergence measure. We discuss to what extent the mapping from nominal weights to main effects can be inverted. This analysis is applied to six composite indicators, including the human development index and two popular league tables of university performance. It is found that in many cases the declared importance of single indicators and their main effect are very different, and that the data correlation structure often prevents developers from obtaining the stated importance, even when modifying the nominal weights in the set of non-negative numbers with unit sum.

Keywords: Composite indicators; Linear aggregation; Modelling; Pearson's correlation ratio; Weights

1. Introduction

In social sciences, composite indicators aggregate individual variables with the aim of capturing relevant, possibly latent, dimensions of reality such as a country's competitiveness (World Economic Forum, 2010), the quality of its governance (Agrast *et al.*, 2010), the freedom of its press (Reporters Sans Frontières, 2011; Freedom House, 2011) or the efficiency of its universities or school system (Leckie and Goldstein, 2009). These measures have been termed 'pragmatic' (see Hand (2009), pages 12–13), in that they answer a practical need to rate individual units (such as countries, universities, hospitals or teachers) for some assigned purpose.

Composite indicators (which are also referred to here as indices) have been increasingly adopted by many institutions, both for specific purposes (such as to determine eligibility for borrowing from international loan programmes) and for providing a measurement basis for shaping broad policy debates, in particular in the public sector (Bird *et al.*, 2005). As a result, public interest in composite indicators has enjoyed a fivefold increase over the period 2005–2010: a search for 'composite indicators' on Google Scholar gave 992 matches in October 2005 and 5340 at the time of the first version of this paper (December 2010).

Composite indicators are fraught with normative assumptions in variable selection and weighting. Here 'normative' is understood to be 'related to and dependent on a system of

Address for correspondence: Paolo Paruolo, Department of Economics, University of Insubria, Via Monte Generoso 71, 21100 Varese, Italy.
E-mail: paolo.paruolo@uninsubria.it

 0964–1998/13/176609

norms and values'. For example, the proponents of the human development index (HDI) advocate replacing gross domestic product (GDP) *per capita* as a measure of the progress of societies with a combination of

(a) GDP *per capita*,
(b) education and
(c) life expectancy;

see Ravallion (2010). Both the selection of these three specific dimensions and the choice of building the index by giving these dimensions equal importance are normative; see Stiglitz *et al.* (2009), page 65. Composite indicators are thus often the subject of controversy; see Saltelli (2007) and Hendrik *et al.* (2008).

The statistical analysis of composite indicators is essential to prevent media and stakeholders from taking them at face value (see the recommendations in Organisation for Economic Co-operation and Development (2008)), possibly leading to questionable choices of policy. For example, a policy maker might think of merging higher education institutions just because the most popular league table of universities puts a prize on larger universities; see Saisana *et al.* (2011).

Most existing composite indicators are linear, i.e. weighted arithmetic averages (Organisation for Economic Co-operation and Development, 2008). Linear aggregation rules have been criticized because weaknesses in some dimensions are compensated by strengths in other dimensions; this characteristic is called 'compensatory'. Non-compensatory and non-linear aggregate ranking rules have been advocated by the literature on multicriteria decision making; see for example Billaut *et al.* (2010), Munda (2008), Munda and Nardo (2009) and Balinski and Laraki (2010). In this paper we concentrate on linear aggregation, because of its widespread use.

We address the issue of measuring variable importance in existing composite indicators. As illustrated by a motivating example at the end of this section, *nominal weights are not a measure of variable importance*, although weights are assigned to reflect some stated target importance, and they are communicated as such. In linear aggregation, the ratio of two nominal weights gives the rate of substitutability between the two individual variables (see Boyssou *et al.* (2006), chapter 4, or Decancq and Lugo (2010)) and hence can be used to reveal the target relative importance of individual indicators. This target importance can then be compared with *ex post* measures of variables' importance, such as the one that is presented in this paper.

We propose to measure the importance of a given variable via Karl Pearson's 'correlation ratio', which is widely applied in global sensitivity analysis as a first-order sensitivity measure; we call this measure the 'main effect'. Main effects represent the expected relative variance reduction obtained in the output (the index) if a given input variable could be fixed (Saltelli and Tarantola (2002); see Section 3.1). They are based on the statistical modelling of the relationship between the variable and the index.

This statistical modelling can be parametric or non-parametric; we compare a linear and a non-parametric alternative based on local linear kernel smoothing. We apply the main effects approach to six composite indicators, including the HDI and two popular league tables of university performance. We find that, in some cases, a linear model can give a reasonable estimate of the main effects, but in other cases the non-parametric fit must be preferred. Further, we find that *nominal weights hardly ever coincide with main effects*. We propose to summarize this deviation in a discrepancy statistic, which can be used by index developers and users alike to gauge the gap between the effective and the target importance of each variable.

We also pose the question of whether the target importance that is stated by the developers is actually attainable by appropriate choice of nominal weights; we call this the 'inverse problem'.

We find that in most instances the correlation structure prevents developers from obtaining the stated importance by changing the nominal weights within the set of non-negative numbers with sum equal to 1. These findings may offer a useful insight to users and critics of an index, and a stimulus to its developers to try alternative, possibly non-compensatory, aggregation strategies.

Our proposed measure of importance is also in line with current practice in sensitivity analysis. Recently, some of the present authors have proposed a global sensitivity analysis approach to test the robustness of a composite indicator (see Saisana *et al.* (2005, 2011)); this approach performs an error propagation analysis of all sources of uncertainty which can affect the construction of a composite indicator. This analysis might be called 'invasive' in that it demands all sources of uncertainty to be modelled explicitly, e.g. by assuming alternative methods to impute missing values, different weights and different aggregation strategies; the method may also test the effect of including or excluding individual variables from the index.

In contrast, the approach that is suggested in this paper is non-invasive, because it does not require explicit modelling of uncertainties. The measure proposed also requires minimal assumptions, in the sense that it exists whenever second moments exist. Moreover, it takes the data correlation structure into account. When this analysis is performed by the developers themselves, it adds to the understanding—and ultimately to the quality—of the index. When performed *ex post* by a third party on an already developed index, this procedure may reveal unnoticed features of the composite indicator.

The paper is organized as follows: the rest of Section 1 reports the motivating example and discusses related work. Section 2 describes linear composite indicators. Section 3 defines the main effects and discusses their estimation. It also defines a discrepancy statistic between main effects and nominal weights. Finally it discusses the inversion of the map from nominal weights to main effects. Section 4 presents detailed results for six indices: the 2009 HDI, the academic ranking of world universities (ARWU) (Center for World-Class Universities, 2008) by Shanghai's Jiao Tong University, the university ranking by the Times Higher Education Supplement (THES) (2008), the 2010 HDI, the index of African governance (IAG) and the sustainable society index (SSI). Section 5 contains a discussion and conclusions. A solution to the inverse problem is reported in Appendix A.

1.1. Motivating example

In weighted arithmetic averages, nominal weights are communicated by developers and perceived by users as a form of judgement of the relative importance of the different variables, including the case of equal weights where all variables are assumed to be equally important. When using 'budget allocation', a strategy to assign weights, experts are given a number of tokens, say 100, and asked to apportion them to the variables composing the index, assigning more tokens to more important variables. This is a vivid example of how weights are perceived and used as measures of importance. However, the relative importance of variables depends on the characteristics of their distribution (after normalization) as well as their correlation structure, as we illustrate with the following example. This gives rise to a paradox, of weights being perceived by users as reflecting the importance of a variable, where this perception can be grossly off the mark.

Consider a university Dean who is asked to evaluate the performance of faculty members, giving equal importance to indicators of publications x_1, of teaching x_2 and of office hours and administrative work x_3. Hence she considers an equally weighted index $y = \frac{1}{3}(x_1 + x_2 + x_3)$, and she employs $R_i^2 := \mathrm{corr}^2(y, x_i)$ to measure the association between the index y and each of the x-variables *ex post*.

We consider two different situations, which illustrate the influence of variances and of correlations of the x-variables on the performance of faculty members. In both situations, we let the variables x_1, x_2 and x_3 be jointly normally distributed with mean 0. First assume that the variance of x_1 is equal to 7 whereas x_2 and x_3 have unit variances, and that the x_j-variables are uncorrelated; the value 7 is chosen here to make the variance of y equal to 1. We then find

$$R_1^2 = 7/9 \approx 0.778,$$
$$R_2^2 = R_3^2 = 1/63 \approx 0.016,$$

which implies that the importance (as measured by R_i^2) of the variables x_2 and x_3 relative to x_1 is equal to $1/49 \approx 0.020$. This shows how variances can greatly affect this measure of importance. We conclude that the Dean needs to do something about the indicators' variances before computing the index.

Changing the weights from $\frac{1}{3}$ to $1/(c\sqrt{\sigma_{ii}})$, where $c := \Sigma_{i=1}^{3} 1/\sqrt{\sigma_{ii}}$ and σ_{ii} is the variance of x_i, would compensate for unequal variances; this corresponds to standardizing indicators before aggregation. In current practice, composite indicators builders prefer to normalize indicators before aggregation, for instance by dividing by the highest score. Going back to the Dean's example, the yearly number of administration hours can be divided by the total number of hours within a year, delivering x_3 as the *fraction* of administration hours. We remark that, in general, normalized scores present different variances.

Consider next the situation where x_1, x_2 and x_3 are standardized, i.e. all have unit variances. Assume also that the correlations $\rho_{ij} := \mathrm{corr}(x_i, x_j)$ are all equal to 0, except $\rho_{23} = \rho_{32} > 0$. Simple algebra shows that

$$R_1^2 = \frac{1}{3 + 2\rho_{23}},$$
$$R_2^2 = R_3^2 = \frac{(1+\rho_{23})^2}{3 + 2\rho_{23}},$$
$$\frac{R_1^2}{R_2^2} = \frac{1}{(1+\rho_{23})^2},$$

i.e. that the importance of indicators x_2 and x_3 is the same; this is a general property of standardized indicators. Note that the importance of indicators x_2 and x_3 is greater than that of x_1, because $\rho_{23} > 0$. Taking for instance $\rho_{23} = 0.7$, we find

$$R_1^2 = 5/22 \approx 0.227,$$
$$R_2^2 = R_3^2 = 289/440 \approx 0.657,$$
$$R_1^2/R_2^2 = 100/289 \approx 0.346.$$

One may imagine a faculty member looking at the relative importance of x_1 with respect to x_2, complaining that research has become dispensable, because—although the index's formula seems to suggest that all variables are equally important—in fact *teaching is valued more than publications by a factor of* 3. In this second situation, even if the Dean has standardized the variables measuring publications x_1, teaching x_2 and administration x_3, the last two have a higher influence on the faculty performance indicator y due to their correlation.

This example describes different situations which generate the paradox. The occurrence of different variances is one such situation; this is a problem also in practice, because usually individual indicators are normalized to be between 0 and 1 or 0 and 100, and hence they have

different variances in general. Also when correcting for different variances by using standardized indicators, however, the paradox can be generated by correlations. This is of practical concern as well, because different individual indicators are usually correlated.

The paradox that was illustrated by the preceding example equally applies when the index's architecture is made of pillars, each pillar aggregating a subset of variables. A hypothetical sustainability index could have environmental, economic, social and institutional pillars, and equal weights for these four pillars would flag the developers' belief that these dimensions share the same importance. Still one of the four pillars with a weighting in principle of 25% could contribute little or nothing to the index, e.g. because the variance of the pillar is comparatively small and/or the pillar is not correlated to the remaining three. A case-study of this nature is discussed later in the present work.

1.2. Related work

The connection of the present paper with global sensitivity analysis has been discussed above. A related approach to measure variable importance in linear aggregations is the one of 'effective weights', which was introduced in the psychometric literature by Stanley and Wang (1968) and Wang and Stanley (1970). The effective weight of a variable x_i is defined as the covariance between $w_i x_i$ and the composite indicator $y = \Sigma_{i=1}^{k} w_i x_i$ divided by its variance, i.e. $\varepsilon_i := \mathrm{cov}(y, w_i x_i)/V(y)$. The same approach has been employed in recent literature in global sensitivity analysis; see for example Li *et al*. (2010).

Effective weights ε_i are, however, not necessarily positive, and hence they make an improper apportioning of the variance $V(y)$: ε_i cannot be interpreted as a 'bit' of variance. In contrast, the measure of importance S_i that is proposed in this paper (i.e. Pearson's correlation ratio) is always positive and can be interpreted as the fractional reduction in the variance of the index that could be achieved (on average) if variable x_i could be fixed. S_i also fits into an analysis-of-variance decomposition framework; see Saltelli (2002) for a discussion.

Moreover, effective weights assume that the dependence structure of the variables x_i is fully captured by their covariance structure, as in linear regression. As we show in what follows, the relationship between the index and its components may be non-linear, and the measure of importance that is proposed in this paper extends to this case as well. The case-studies that are reported in Section 4 show that non-linearity is often the rule rather than the exception. In the case of a linear relationship between y and x_i, our measure S_i reduces to R_i^2, the square of $\mathrm{corr}(y, x_i)$, used in the example above; hence in this case, the present approach leads to a simple transformation of the effective weights.

For some indices, such as the product market regulation index (see Nicoletti *et al*. (2000)), principal component analysis (PCA) has been used to select aggregation weights. PCA chooses weights that maximize (or minimize) the variance of the index, and hence weights do not reflect the normative aspects of the definition of the index. Consequently, weights are difficult to interpret and to communicate, and as a result the use of PCA in this context is not widespread. The same product market regulation index moved from the use of PCA to a simpler and more transparent technique for linear aggregation after a statistical analysis of the implications of such a change (Nardo, 2009).

2. Weights and importance

Consider the case of a composite indicator y calculated as a weighted arithmetic average of k variables x_i,

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

where x_{ji} is the normalized score of individual j (e.g. country) based on the value X_{ji} of variable $X_{\cdot i}$, $i = 1, \ldots, k$, and w_i is the nominal weight assigned to variable $X_{\cdot i}$. The most common approach is to normalize original variables (see Bandura (2008)), by the min–max normalization method

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

where $X_{\max,i}$ and $X_{\min,i}$ are the upper and lower values respectively for the variable $X_{\cdot i}$; in this case all scores x_{ji} vary in [0, 1]. Here we indicate transformation (2) as 'normalization'; the normalized variables in equation (2) are denoted as $x_{\cdot i}$. We let $\mu_i := E(x_{ji})$ and $\sigma_{ii} = V(x_{ji})$ indicate their expectation and variance respectively. In what follows, we replace $X_{\cdot i}$ and $x_{\cdot i}$ by X_i and x_i respectively, unless needed for clarity.

Observe that the normalization (2) implies a fixed scale of the individual indicators; this is useful for instance for comparability in repeated waves of the same index. However, normalization does not imply any standardization of different $x_{\cdot i}$-variables, which hence have different means μ_i and variances σ_{ii} in general.

A popular alternative to the min–max normalization (2) is given by standardization

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

where $E(X_{ji})$ and $V(X_{ji})$ are the mean and variances of the original variables $X_{\cdot i}$. When standardized, all x_i have the same mean and variance, $\mu_i = 0$ and $\sigma_{ii} = 1$ for all i, removing one source of heterogeneity among variables. However, standardization does not affect the correlation structure of the variables X_i (or x_i). Both transformations (2) and (3) are invariant to the choice of unit of measurement of X_i; see Hand (2009), chapter 1.

Although standardization may appear a better approach than normalization, statistically, there are advantages and disadvantages of both. For example standardization may be expected not to work so well when the distribution is very skewed or long tailed. Moreover it does not enhance comparison across different waves of the same aggregate indicator over the years, if the mean and variances that are used in equation (3) change over time. Also one cannot achieve *both* standardization and normalization at the same time through a linear transformation of X_i. This implies that index developers suffer the unwanted disadvantages of the transformation chosen.

Whatever the transformation, in what follows we denote the column vector of scores of unit j as $\mathbf{x}_j := (x_{j1}, \ldots, x_{jk})'$ and indicate by $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_k)'$ and $\boldsymbol{\Sigma} := (\sigma_{it})_{i,t=1}^{k}$ the corresponding vector of means and the implied variance–covariance matrix. The weight w_i that is attached to each variable x_i in the aggregate is meant to appreciate the importance of that variable with respect to the concept being measured. The vector of weights $\mathbf{w} := (w_1, \ldots, w_k)'$ is selected by developers on the basis of different strategies, be those statistical, such as PCA, or based on expert evaluation, such as an analytic hierarchy process; see Saaty (1980, 1987).

In what follows we indicate by ζ_{il}^2 the target relative importance of indicators i and l. When this is not explicitly stated, the ratios w_i/w_l can be taken to be the 'revealed target relative importance'. In fact w_i/w_l is a measure of 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 (2010).

For simplicity of notation and without loss of generality, we assume that the maximal weight is assigned to indicator 1, i.e. that $w_1 \geqslant w_i$ for $i = 2, \ldots, k$, and we consider $\zeta_i^2 := \zeta_{i1}^2$.

The previous discussion applies to pillars as well as to individual variables, where a pillar is defined as an aggregated subset of variables, identified by the developers as representing a salient—possibly latent, or normative—dimension of the composite indicator.

3. Measuring importance

3.1. Measures of importance

In this paper we propose a variance-based measure of importance. We note that

$$\begin{aligned} E(y) &= \mathbf{w}'\boldsymbol{\mu}, \\ V(y) &= \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}, \end{aligned} \tag{4}$$

where, if equation (3) is used, $E(y) = 0$ and the diagonal elements of $\boldsymbol{\Sigma}$ are equal to 1; here we have dropped the subscript j in y_j for conciseness. In what follows, we focus attention on the variance term.

Following Pearson (1905), we consider the question 'what would be the average variance of y, if variable x_i were held fixed?'. This question leads us to consider

$$E_{x_i}\{V_{\mathbf{x}_{\sim i}}(y|x_i)\},$$

where $\mathbf{x}_{\sim i}$ is defined as the vector containing all the variables in $\mathbf{x}$ except variable x_i. Owing to the well-known identity

$$V_{x_i}\{E_{\mathbf{x}_{\sim i}}(y|x_i)\} + E_{x_i}\{V_{\mathbf{x}_{\sim i}}(y|x_i)\} = V(y)$$

we can define the ratio of $V_{x_i}\{E_{\mathbf{x}_{\sim i}}(y|x_i)\}$ to $V(y)$ as a measure of the relative reduction in variance of the composite indicator to be expected by fixing a variable, i.e.

$$S_i \equiv \eta_i^2 := \frac{V_{x_i}\{E_{\mathbf{x}_{\sim i}}(y|x_i)\}}{V(y)}. \tag{5}$$

The notation S_i reflects the use of this measure as a first-order sensitivity measure (also termed 'main effect') in sensitivity analysis; see Saltelli and Tarantola (2002). The notation η_i^2 reflects the original notation that was used in Pearson (1905); he called it 'correlation ratio η^2'.

The conditional expectation $E_{\mathbf{x}_{\sim i}}(y|x_i)$ in the numerator of expression (5) can be any non-linear function of x_i; in fact

$$f_i(x_i) := E_{\mathbf{x}_{\sim i}}(y|x_i) = w_i x_i + \sum_{l=1, l\neq i}^{k} w_l \, E_{\mathbf{x}_{\sim i}}(x_l|x_i),$$

where the latter conditional expectations may be linear or non-linear in x_i. For the connection of $f_i(x_i)$ to global sensitivity analysis see Saltelli *et al.* (2008).

In the special case of $f_i(x_i)$ linear in x_i, we find that S_i reduces to R_i^2, where R_i is the product moment correlation coefficient of the regression of y on x_i. In fact, it is well known that when f_i is linear, i.e. $f_i(x_i) = \alpha_i + \beta_i x_i$, it coincides with the L_2-projection of y on x_i, which implies that $\beta_i = \mathrm{cov}(y, x_i)/\sigma_{ii}$; see for example Wooldridge (2010). Hence S_i has the form $S_i = V_{x_i}(\beta_i x_i + \alpha_i)/V(y)$ and we find that $S_i = \beta_i^2 \sigma_{ii}/V(y) = \mathrm{cov}^2(y, x_i)/\{\sigma_{ii} V(y)\} = R_i^2$.

A further special case corresponds to f_i linear and $\mathbf{x}$ made of uncorrelated components. We find that $\mathrm{cov}(y, x_i) = \Sigma_{t=1}^k w_t \sigma_{ti}$ and $V(y) = \Sigma_{t=1}^k w_t^2 \sigma_{tt}$ so $S_i = w_i^2 \sigma_{ii}/\Sigma_{t=1}^k w_t^2 \sigma_{tt}$. The main difference between the uncorrelated and the correlated case is that in the former $\Sigma_{i=1}^k S_i = 1$ because

$S_i = w_i^2\sigma_{ii}/\Sigma_{h=1}^k w_h^2\sigma_{hh}$, whereas for the latter $\Sigma_{i=1}^k S_i$ might exceed 1; see for example Saltelli and Tarantolla (2002). We note that in general S_i can still be high also when R_i^2 is low, e.g. in the case of a non-monotonic U-shaped relationship for $f_i(x_i)$. Hence in general $f_i(x_i)$ needs to be estimated in a non-parametric way; see Section 3.2.

As these special cases illustrate, S_i is a quadratic measure in terms of the weights w_j for linear aggregation schemes (1); this follows from its definition as a variance-based measure. The main effect S_i is an appealing measure of importance of a variable (be it indicator or pillar) for several reasons.

(a) It offers a precise definition of importance of a variable, i.e. 'the expected fractional reduction in variance of the composite indicator that would be obtained if that variable could be fixed'.
(b) It can be applied when relationships between the index and its components are linear or non-linear. Such non-linearity may be the effect of non-linear aggregation (e.g. Condorcet like; see Munda (2008)) and/or of non-linear relationships between the single variables. It can be used regardless of the degree of correlation between variables. Unlike the Pearson or Spearman correlation coefficients, it is not constrained by assumptions of linearity or monotonicity.
(c) It is not invasive, i.e. no changes are made to the composite indicator or to the correlation structure of the indicators, unlike for example the error propagation analysis that was presented in Saisana *et al.* (2005). Whereas the error propagation can be considered as a stress test of the index, the present approach is a test of its internal coherence.

3.2. Estimating main effects

In this subsection we consider estimating the main effects and focus on the 2009 HDI to illustrate our approach. In Section 4 we describe the six case-studies of our approach in detail.

In sensitivity analysis, the estimation of S_i is an active research field. S_i can be estimated from design points (Sobol', 1993; Saltelli, 2002; Saltelli *et al.*, 2010), Fourier analysis (Tarantola *et al.*, 2006; Plischke, 2010; Xu and Gertner, 2011) or others. Many non-parametric estimators can be used to estimate $f_i(x_i)$, such as state-dependent regression; see Ratto *et al.* (2007) and Ratto and Pagano (2010).

In the present work we employ a non-parametric, local linear, kernel regression to estimate $m(\cdot) := f_i(\cdot)$, and then use it in expression (5) to estimate S_i, replacing the variances in the numerator and denominator with the corresponding sample variances, i.e. using

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

where $\bar{y} := n^{-1}\Sigma_{j=1}^n y_j$, $\bar{m} := n^{-1}\Sigma_{j=1}^n m_j$, $m_j := \hat{m}(x_{ji})$ and $\hat{m}(\cdot)$ is the estimate of $m(\cdot) := f_i(\cdot)$.

Local linear kernel estimators achieve automatic boundary corrections and enjoy some typical optimal properties that are superior to Nadaraya–Watson kernel estimators; see Ruppert and Wand (1994) and reference therein. As a result, local linear kernel smoothers are often considered the standard non-parametric regression method; see for example Bowman and Azzalini (1997).

The local linear non-parametric kernel regression is indexed by a bandwidth parameter h, which is usually held constant across the range of values for x_i. For large h, the local linear non-parametric kernel regression converges to the linear least squares fit. This allows us to interpret $1/h$ as the deviation from linearity; it suggests that we investigate the sensitivity of the

estimation of S_i to variation in the bandwidth parameter h. To make this dependence explicit we write $S_i(h)$ to indicate the value of S_i that is obtained by a local linear kernel regression with bandwidth parameter h. In the application we use a Gaussian kernel.

The choice of the smoothing parameter h can be based either on cross-validation (CV) principles (see Bowman and Azzalini (1997)) or on plug-in choices for the smoothing parameter, such as those proposed in Ruppert *et al.* (1995). We describe these approaches in turn, starting with CV. Let $\hat{m}(x)$ indicate the local linear non-parametric kernel estimate for $f_i(x)$ at $x_i = x$ based on all n observations, and let $\hat{m}_{-j}(x)$ be the same applied to all data points except for that with index j; then the least squares CV criterion for variable x_i is defined as

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{j=1}^{n}\{y_j - \hat{m}_{-j}(x_{ji})\}^2.$$

The optimal value for the CV criterion is given by the bandwidth h_{CV} corresponding to the minimum of $\mathrm{CV}(h)$. In practice, a grid $\mathcal{H}$ of possible values for h is considered, and the minimum of the function $\mathrm{CV}(h)$ is found numerically. In the application we chose the grid of h-values as follows: we defined a regular grid of 50 values for $u := \sqrt{h}$ in the range from 0.1 to 5. The values for h were then obtained as $h = a + u^2/b$, for index-specific constants a and b; the resulting set of values in this grid is denoted $\mathcal{H}$ in what follows.

The default values for indices with range from 0 to 10 or 100 were $a = 0.05$ and $b = 1$, so $0.06 < h \leqslant 25.05$; for indices with range from 0 to 1 (namely the 2009 and 2010 HDI), we chose $a = 0.01$ and $b = 25$, so $0.01 < h \leqslant 1.01$. In some cases $\mathrm{CV}(h)$ attains its minimum at the right-hand end of the grid $\mathcal{H}$; this happened for both the ARWU$\{1, 2, 3\}$ and THES$\{4, 6\}$ indices (see Table 1 in Section 4) as well as for the IAG$\{2, 5\}$ and SSI$\{2\}$ indices (see Table 5 in Section 4), where the digits in braces refer to the subscript i of the x_i-variables. In these cases, in practice, a linear regression fit would not be worse than the fit of the local linear kernel estimator, according to the CV criterion.

In the implementation of the CV criterion, when a local linear kernel regression implied a row of the smoothing matrix with numerical 'divisions by 0', we replaced it with a local mean (Nadaraya–Watson) estimator. When also the latter would imply numerical divisions by 0, we replaced the row of the smoothing matrix with a sample leave-one-out mean.

An alternative choice of bandwidth is given by plug-in-rules. One popular choice is given by the 'direct plug-in' (DPI) selector h_{DPI} that was introduced by Ruppert *et al.* (1995), which minimizes the asymptotic mean integrated squared error for the local linear Gaussian kernel smoother, on the basis of the following preliminary estimators. Let $\theta_{rs} := E(m^{(s)}m^{(r)})$, where $m^{(r)}(x)$ is the rth derivative of $m(x)$. The range of x_i is partitioned into N blocks and a quartic is fitted on each block. Using this estimation, an estimate for θ_{24} is found, along with an estimator for the error variance $\sigma^2 := E\{y_j - m(x_{ji})\}^2$. These estimates are then used to obtain a plug-in bandwidth g, which is used in a local cubic fit to estimate θ_{22} and to obtain a different plug-in bandwidth λ. The λ-bandwidth is then used in a final local linear kernel smoother to estimate σ^2, which is fed into the final formula for h_{DPI}, along with the previous estimate of θ_{22}. The choice of N, the number of blocks, is obtained minimizing Mallows's C_p-criterion over the set $\{1, 2, \ldots, N_{\max}\}$, where $N_{\max} = \max\{\min(\lfloor n/20 \rfloor, N^*), 1\}$.

In the application we chose $N^* = 5$ as suggested by Ruppert *et al.* (1995); in case of numerical instabilities, we decreased N^* to 4. Moreover we performed an α-trimming in the estimation of θ_{24} and θ_{22} with $\alpha = 0.05$. Because the choice of bandwidth can be affected by values at the end of the x-range, we considered only pairs of observations for which $x > 0$ in the choice of bandwidth, for both the CV criterion and the DPI criterion.

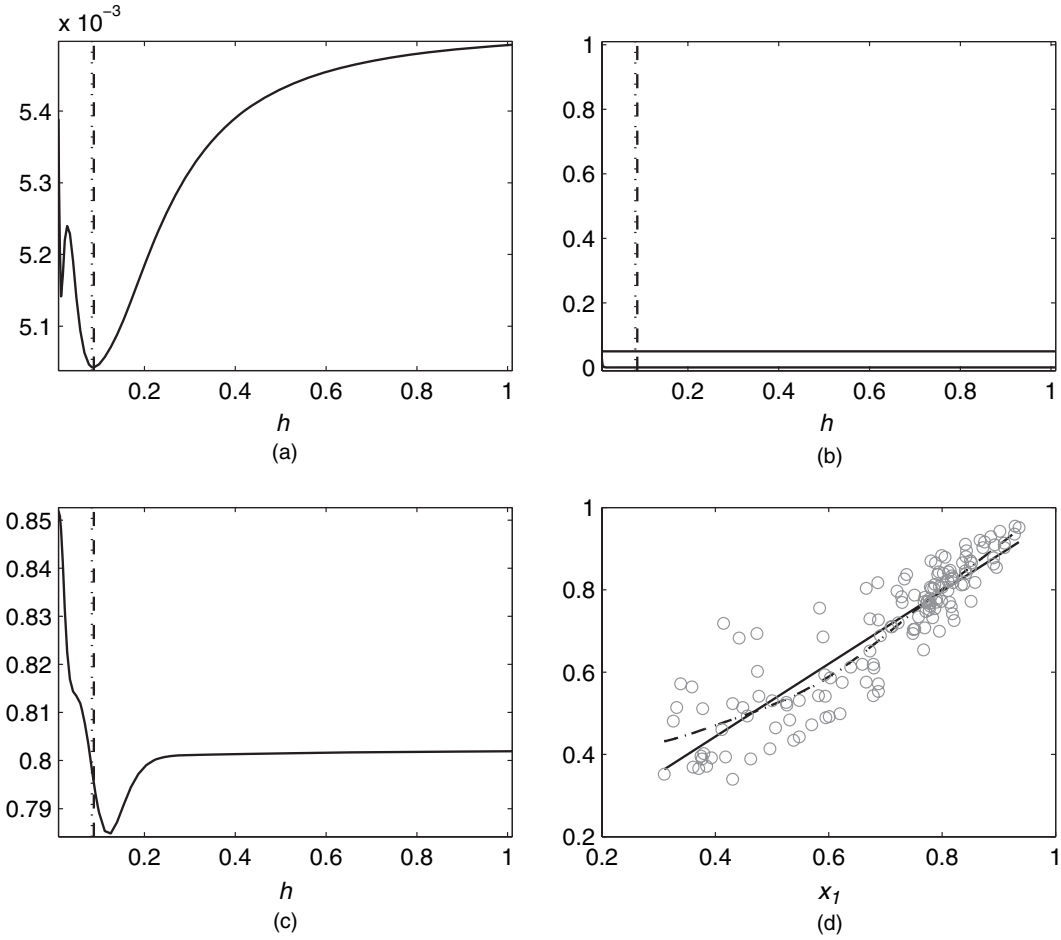

Fig. 1. 2009 HDI y and life expectancy x_1: (a) CV criterion as a function of the smoothing parameter h (⋮, h_{DPI}; ⋮, h_{CV}); (b) linearity test p-value as a function of h (⋮, h_{DPI}; ⋮, h_{CV}); (c) main effects S_i as a function of h (⋮, h_{DPI}; ⋮, h_{CV}); (d) cross-plot of y *versus* x_1 with a linear fit and local linear fits for $h_{\mathrm{DPI}} = 0.0841$ (·····) and $h_{\mathrm{CV}} = 0.088$ (· — · —)

The resulting choice of bandwidth h_{DPI} was sometimes very close to h_{CV}, as in the case for the 2009 HDI, which is depicted in Figs 1–4, where each figure refers to one of the four x_i-indicators that were used in the construction of the 2009 HDI. Fig. 1 refers to the x_1-indicator (life expectancy) and contains four panels, which report—counterclockwise from Fig. 1(b)—the p-value of the linearity test introduced below, the CV criterion, the S_1-measure and the regression cross-plot. Figs 1(a)–1(c) show functions of the bandwidth parameter h, whereas Fig. 1(d) has the values of x_1 on the horizonal axis. Figs 2–4 have the same format, and refer to indicators x_2, x_3 and x_4.

Tables 1 and 5 in Section 4 report the selected values of h_{CV} and h_{DPI} for the 2009 HDI and for the other five indices, which are described in detail in Section 4. It can be seen that the values of h_{CV} sometimes differed from h_{DPI} by several orders of magnitude.

As in many other contexts, in the estimation of main effects S_i the linear case is a relevant reference model, and we would like to address inference on S_i and on the possible linearity of $f_i(x_i)$ jointly. For this we implemented the test for linearity that is proposed in Bowman and

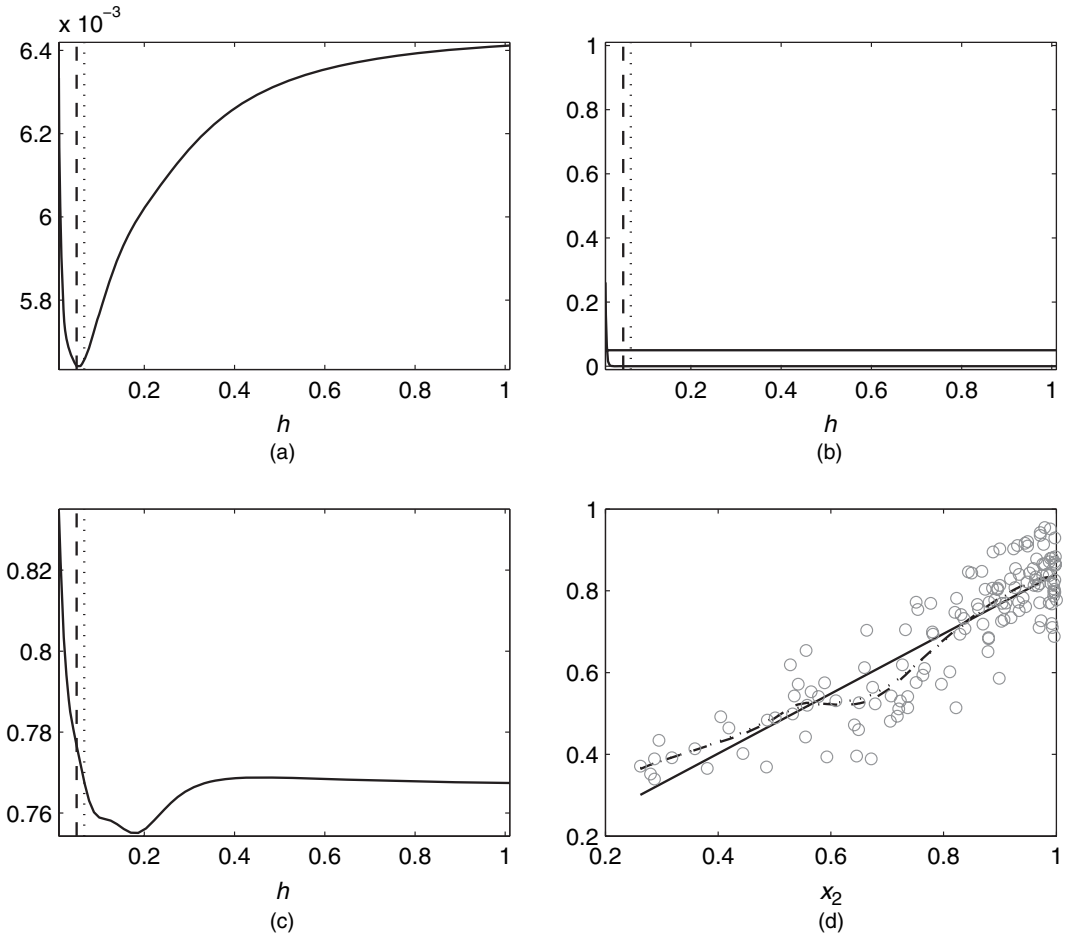

Fig. 2. 2009 HDI y and adult literacy x_2: (a) CV criterion as a function of the smoothing parameter h (⋮, h_{DPI}; ⋮, h_{CV}); (b) linearity test p-value as a function of h (⋮, h_{DPI}; ⋮, h_{CV}); (c) main effects S_i as a function of h (⋮, h_{DPI}; ⋮, h_{CV}); (d) cross-plot of y *versus* x_2 with a linear fit and local linear fits for $h_{\mathrm{DPI}} = 0.0666$ (·····) and $h_{\mathrm{CV}} = 0.05$ (· — · —)

Azzalini (1997), chapter 5. The fit of the linear kernel smoother can be represented as $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, where the matrix $\mathbf{S}$ depends on all values x_{ji}, $j = 1, \ldots, n$. A test of linearity can be based on the F-statistic, $F := (\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1$, that compares the residual sum of squares under the linearity assumption RSS_0 with the one corresponding to the local linear kernel smoother RSS_1. Letting F_{obs} indicate the value of the statistic, the p-value of the test is computed as the probability that $\mathbf{z}'\mathbf{C}\mathbf{z} > 0$ where $\mathbf{z}$ is a vector of independent standard Gaussian random variables and $\mathbf{C} := \mathbf{M}(\mathbf{I} - (1 + F_{\mathrm{obs}})\mathbf{A})\mathbf{M}$ with $\mathbf{A} = (\mathbf{I} - \mathbf{S})'(\mathbf{I} - \mathbf{S})$, $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{X}$ equal to the linear regression design matrix, with first column equal to the constant vector and the second column equal to the values of x_{ji}, $j = 1, \ldots, n$.

Bowman and Azzalini (1997) suggested approximating the quantiles of the quadratic form with the distribution of $a\chi^2_b + c$, where a, b and c are obtained by matching moments of the quadratic form and the $(a\chi^2_b + c)$-distribution; here χ^2_b represents a χ^2-distribution with b degrees of freedom. We implemented this approximation; Figs 1(b), 2(b), 3(b) and 4(b) report the resulting p-values of the test as a function of h for the 2009 HDI. It can be seen that for some x_i-variable

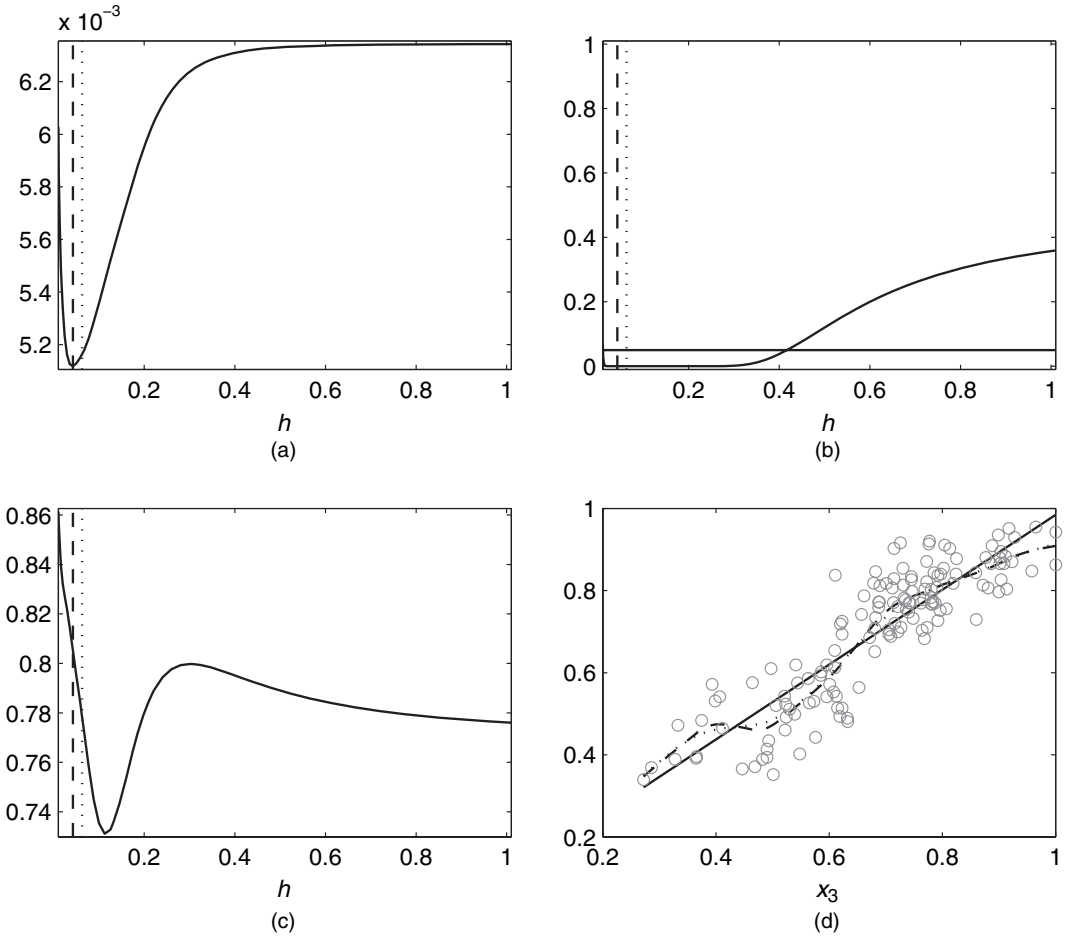

Fig. 3. 2009 HDI y and enrolment in education x_3: (a) CV criterion as a function of the smoothing parameter h (⋮, h_{DPI}; ⋮, h_{CV}); (b) linearity test p-value as a function of h (⋮, h_{DPI}; ⋮, h_{CV}); (c) main effects S_i as a function of h (⋮, h_{DPI}; ⋮, h_{CV}); (d) cross-plot of y *versus* x_3 with a linear fit and local linear fits for $h_{\mathrm{DPI}} = 0.0631$ (⋯⋯) and $h_{\mathrm{CV}} = 0.0424$ (· — · —)

the test rejects the linearity hypothesis for all values of h in the grid $\mathcal{H}$, and for some other pairs the test rejects only for a subset of $\mathcal{H}$. In a few other pairs, the test never rejects for all $h \in \mathcal{H}$. Results for the linearity test are reported in Tables 1 and 5 for selected values of h, for the 2009 HDI and for five other indices, which are described in detail in Section 4.

To show sensitivity of the main effects S_i to the smoothing parameter h, we also computed the $S_i(h)$ index as a function of h. We also recorded the minimum and maximum values obtained for $S_i(h)$ varying h in $\mathcal{H}$; we denote these values $S_{i,\min}$ and $S_{i,\max}$. We report the plot of $S_i(h)$ as a function of h in Figs 1(c), 2(c), 3(c) and 4(c).

3.3. Comparing weights and main effects

In this section we compare revealed or target relative importance measures ζ_i^2 with the relative main effects S_i/S_1. First note that, in the independent case,

$$S_i = w_i^2 \sigma_{ii} \Big/ \sum_{h=1}^{k} w_h^2 \sigma_{hh},$$

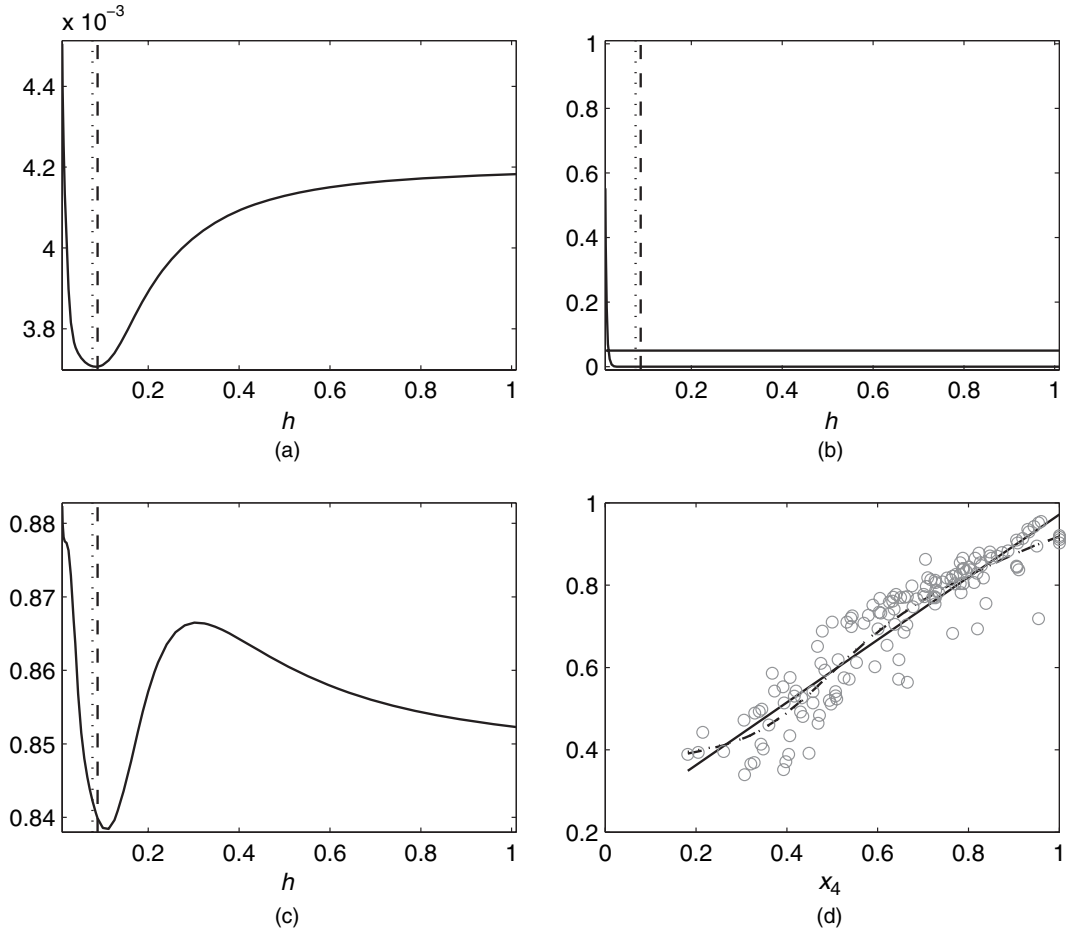

Fig. 4. 2009 HDI y and GDP *per capita* x_4: (a) CV criterion as a function of the smoothing parameter h (⋮, h_{DPI}; ⋮, h_{CV}); (b) linearity test p-value as a function of h (⋮, h_{DPI}; ⋮, h_{CV}); (c) main effects S_i as a function of h (⋮, h_{DPI}; ⋮, h_{CV}); (d) cross-plot of y *versus* x_4 with linear fit and local linear fits for $h_{\mathrm{DPI}} = 0.0771$ (·····) and $h_{\mathrm{CV}} = 0.0884$ (· — · —)

so $S_i/S_1 = w_i^2\sigma_{ii}/w_1^2\sigma_{11}$. When the x_i-variables are standardized, all $\sigma_{ii} = 1$ and hence $S_i/S_1 = w_i^2/w_1^2$. The relative main effects S_i/S_1 do not reduce to $\zeta_i^2 = w_i/w_1$, except in the homoscedastic case ($\sigma_{ii} = \sigma_{11}$) when the nominal weights are equal ($w_i = w_1$), so $w_i^2\sigma_{ii}/w_1^2\sigma_{11} = w_i^2/w_1^2 = 1 = w_i/w_1$. In the general case, S_i depends on $\mathbf{w}$ and $\boldsymbol{\Sigma}$ in a more complicated way, and hence there is no reason, *a priori*, to expect S_i/S_1 to coincide with ζ_i^2.

One can compare how the effective relative importance S_i/S_1 deviates from the (revealed) target relative importance ζ_i^2; for this we define the maximal discrepancy statistic d_m as

$$d_m = \max_{i\in\{2,\ldots,k\}} |\zeta_i^2 - S_i/S_1|. \tag{6}$$

In the case of revealed target relative importance, recall that w_1 is assumed to be the highest nominal weight $w_{\max}$. In the case when more than one variable has maximum weight equal to $w_{\max}$, we selected as reference variable the variable with maximum value for $S_l(h_{l,\mathrm{DPI}})$ with $l \in \{1,\ldots,k\}$, i.e. $l = \arg\max_{i\in\{1,\ldots,k\}} S_i(h_{i,\mathrm{DPI}})$ where $h_{i,\mathrm{DPI}}$ is the DPI bandwidth choice for indicator i.

The higher the value of d_m, the more discrepancy there is between relative target importance and the corresponding relative main effects. In d_m we have chosen to capture the discrepancy by focusing on the maximal deviation; alternatively we can consider any absolute power mean, f-divergence function or distance between the (unnormalized) distributions $\{\zeta_i^2\}$ and $\{S_i/S_1\}$. For simplicity, in what follows we indicate these distributions used in the comparison as $\{\zeta_i^2\}$ and $\{S_i\}$.

Because d_m depends on the choice of bandwidth parameters h in the estimation of $S_i(h)$, $i=1,\ldots,k$, we also calculated bounds on the variation of d_m obtained by varying h. Specifically, we computed d_m comparing $\{\zeta_i^2\}$ with $\{S_{i,l_i}\}$ choosing l_i as either equal to the minimum or maximum, considering all possible combinations. For instance, with $k=2$, we considered $\{S_{1,\min}, S_{2,\min}\}$, $\{S_{1,\min}, S_{2,\max}\}$, $\{S_{1,\max}, S_{2,\min}\}$ and $\{S_{1,\max}, S_{2,\max}\}$. Within the distribution of values of d_m that were obtained in this way, we recorded the minimum and the maximum, denoted as $d_{m,\min}$ and $d_{m,\max}$. Table 3 in Section 4 reports the d_m for h equal to h_{DPI} and h_{CV} and in the linear case, along with the values $d_{m,\min}$ and $d_{m,\max}$, which provide a measure of sensitivity of d_m with respect to the choice of bandwidth h.

To compare the S_i-values with the weights w_i graphically, in Fig. 5 in Section 4 we rescale the S_i-values to have sum equal to 1, considering $S_i^* := S_i/c$ with $c := \Sigma_{t=1}^k S_t$, which we call 'normalized S_i'. To visualize bounds for S_i^*, we plot bars with end points equal to $S_{i,\min}/c$ and $S_{i,\max}/c$; these bars inform on the sensitivity of S_i^* with respect to the variation of the bandwidth parameter h.

3.4. Reverse engineering the weights

This section discusses when it is possible to find nominal weights w_i that imply predetermined given values z_i^2 for the relative main effects S_i/S_1; here we indicate the target relative importance z_i^2 to differentiate it from ζ_i^2 of the previous sections. This reverse engineering exercise can help developers of composite indicators to anticipate criticism by enquiring whether the stated relative importance of pillars or indicators is actually attainable.

For the purpose of this inversion, we consider the case of $f_i(x_i)$ linear in x_i; in this case S_i coincides with R_i^2, the square of Pearson's product moment correlation coefficients between y and x_i. The linear case can be seen as a first-order approximation to the non-linear general case; this choice is motivated by the fact that we can find an exact solution to the inversion problem of the map from w_i to R_i^2/R_1^2 when we allow weights w_i also to be negative. One expects that the reverse engineering formula in the linear case will be indicative of the formula based on a non-linear approach, where the latter would be computationally more demanding.

We wish to find a value $\mathbf{w}^* := (w_1^*, \ldots, w_k^*)'$ for the vector of nominal weights $\mathbf{w} := (w_1, \ldots, w_k)'$ such that R_i^2/R_1^2 equals preselected target values z_i^2, for $i=1,\ldots,k$. We call this the 'inverse problem'. The weights w_i^* are chosen to sum to 1, but they are allowed also to be negative; this choice makes the inverse problem solvable, and in Appendix A we show that it has a unique solution, given by

$$\mathbf{w}^* = \frac{1}{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{g}}\boldsymbol{\Sigma}^{-1}\mathbf{g}, \tag{7}$$

where $\mathbf{g}$ is a vector with ith entry equal to $g_i := z_i\sqrt{(\sigma_{ii}/\sigma_{11})}$ and $\mathbf{1}$ is a k-vector of 1s.

Because the solution to this inverse problem is unique, if some of the weights w_j^* in equation (7) are negative, it means that a solution to the inverse problem with all positive weights does not exist and hence the targets z_i^2 are not attainable, owing to the data covariance structure. This can help designers to reformulate their targets to make them attainable, and the stakeholders

involved in the use of the composite indicator to evaluate whether the individual indicators can have the stated importance by an appropriate choice of weights.

4. Case-studies

In this section we apply the statistical analysis that was described in Section 3 to the six composite indicators. In Section 4.1 we consider the three indices for which aggregation was performed at indicator level and in Section 4.2 we consider the three indices for which aggregation was performed at the pillar level.

4.1. Importance at the indicator level

We consider the HDI and two well-known composite indicators of university performance: the ARWU of Shanghai's Jiao Tong University and the indicator that is associated with the UK's THES.

4.1.1. University ranking

The 2008 ARWU (see Center for World-Class Universities (2008)) summarizes quality of education, quality of faculty, research output and academic performance of world universities by using six indicators: the number of alumni of an institution having won Nobel Prizes or Fields Medals (weight 10%), the number of Nobel or Fields laureates among the staff of an institution (weight 20%), the number of highly cited researchers (weight 20%), the number of articles published in *Nature* or *Science*, the expanded *Science Citation Index* and *Social Sciences Citation Index* (weight 40%) and finally academic performance measured as the weighted average of these five indicators divided by the number of full-time equivalent academic staff (weight 10%). The raw data are normalized by assigning to the best performing institution a score of 100 and all other institutions receiving a score relative to the leader. The ARWU score is a weighted average of the six normalized indicators, which is finally rescaled to a maximum of 100. The six indicators have moderate to strong correlations in the range from 0.48 to 0.87 and an average bivariate correlation of 0.68.

The 2008 THES (see Times Higher Education Supplement (2008)) summarizes university features related to research quality, graduate employability, international orientation and teaching quality by using six indicators: the opinion of academics on which institutions they consider to be the best in the relevant field of expertise (weight 40%), the number of papers published and citations received by research staff (weight 20%), the opinion of employers about the universities from which they would prefer to recruit graduates (weight 10%), the percentage of overseas staff at the university (weight 5%), the percentage of overseas students (weight 5%) and finally the ratio between the full-time equivalent faculty and the number of students enrolled at the university (weight 20%). The raw data are standardized. The standardized indicator scores are then scaled by dividing by the best score. The THES score is the weighted average of the six normalized indicators, which is finally rescaled to a maximum of 100. The six indicators have very low to moderate correlations that range from 0.01 to 0.64 and a low average bivariate correlation of 0.24.

Results for the ARWU and THES are given in Tables 1–3. The first two panels of Table 1 provide the bandwidth selection results for the ARWU and THES; the corresponding panels of Table 2 give estimates of the importance measure S_i for various choices of bandwidth. The first two rows in Table 3 give the maximum discrepancy statistic d_m for the ARWU and THES. Finally Figs 5(a) and 5(b) summarize the comparison between target and actual relative impor-

Table 1. Bandwidth choice at indicator level†

	h_{CV}	p_{CV}	h_{DPI}	p_{DPI}	n
2008 ARWU					
Alumni winning Nobel Prize	25.05‡	0.88	3.43	0.71	198
Staff winning Nobel Prize	25.05‡	0.59	3.13	0.27	135
Highly cited research	25.05‡	0.00	1.15	0.00	424
Articles in *Nature* and *Science*	9.05	0.00	1.78	0.00	494
Articles in *Science* and *Social Sciences Citation Index*	2.94	0.00	2.26	0.00	503
Academic performance (size adjusted)	1.74	0.00	2.12	0.00	503
2008 THES					
Academic review	4.46	0.00	1.74	0.00	400
Recruiter review	5.81	0.00	2.62	0.00	400
Teacher/student ratio	4.46	0.07	4.76	0.08	399
Citations per faculty	25.05‡	0.04	2.44	0.20	400
International staff	6.81	0.04	2.97	0.22	398
International students	25.05‡	0.18	4.13	0.65	399
2009 HDI					
Life expectancy	0.09	0.00	0.08	0.00	142
Adult literacy	0.05	0.00	0.07	0.00	142
Enrolment in education	0.04	0.00	0.06	0.00	142
GDP *per capita*	0.09	0.00	0.08	0.00	142

†Bandwidths $h_{i,CV}$ and $h_{i,DPI}$ and corresponding p-values for the linearity test $p_{i,CV}$ and $p_{i,DPI}$; n is the number of observations with $x_{ji} > 0$ used for CV and DPI.
‡Right-hand end of the grid $\mathcal{H}$.

tance of indicators. For the ARWU the main effects S_i are more similar to each other than the nominal weights, i.e. ranging between 0.14 and 0.19 (normalized S_i-values to unit sum; CV estimates) when weights should either be 0.10 or 0.20.

The situation is worse for the THES index, where the combined importance of peer-review-based variables (recruiters and academia) appears larger than stipulated by developers, indirectly supporting the hypothesis of linguistic bias at times addressed to this measure (see for example Saisana *et al*. (2011) for a review). Further for the THES index the 'teachers-to-student ratio', a key variable aimed at capturing the teaching dimension, is much less important than it should be when comparing normalized S_i (0.09; CV estimate) with the nominal weight (0.20).

Overall, there is more discrepancy between the nominal weights that were assigned to the six indicators and their respective main effects in the THES ranking ($d_{m,CV} = 0.42$) than in the ARWU ($d_{m,CV} = 0.36$) CV estimates. Comparing this result with the conclusions in Saisana *et al*. (2011), we can see the value added of the present measure of importance. From Saisana *et al*. (2011) we could not make a judgement about the relative quality of the THES ranking with respect to the ARWU. The main effects that are used here allow us to say that—leaving aside the different normative frameworks about which no statistical inference can be made—the ARWU is statistically more consistent with its declared targets than the THES index.

When considering the sensitivity of d_m-values to the choice of bandwidths h, we can see that the range $[d_{m,\min}, d_{m,\max}]$ is slightly shorter for the ARWU ($[0.26, 0.50]$) than for the THES index ($[0.29, 0.55]$); this implies that the ARWU is slightly less sensitive than the THES index to the choice of bandwidths h. Note, however, that the two ranges overlap, so there are choices of bandwidths h for which the ordering of d_m-values is reversed. This, however, does not happen at the values h_{CV} and h_{DPI}.

Table 2. Main effects at indicator level†

	w_i	$S_{i,\text{lin}}$	$S_{i,\text{CV}}$	$S_{i,\text{DPI}}$	$S_{i,\min}$	$S_{i,\max}$
2008 ARWU						
Alumni winning Nobel Prize	0.10	0.64	0.65	0.67	0.65	0.76
Staff winning Nobel Prize	0.20	0.72	0.72	0.73	0.72	0.80
Highly cited researchers	0.20	0.81	0.85	0.87	0.85	0.90
Articles in *Nature* and *Science*	0.20	0.87	0.88	0.88	0.88	0.94
Articles in *Science* and *Social Sciences Citation Index*	0.20	0.63	0.70	0.70	0.64	0.90
Academic performance (size adjusted)	0.10	0.71	0.76	0.75	0.72	0.88
2008 THES						
Academic review	0.40	0.77	0.81	0.82	0.78	0.85
Recruiter review	0.10	0.45	0.54	0.54	0.46	0.62
Teacher/student ratio	0.20	0.19	0.21	0.20	0.18	0.42
Citations per faculty	0.20	0.38	0.38	0.41	0.38	0.50
International staff	0.05	0.10	0.12	0.12	0.10	0.31
International students	0.05	0.16	0.16	0.17	0.16	0.34
2009 HDI						
Life expectancy	0.33	0.80	0.80	0.80	0.78	0.85
Adult literacy	0.22	0.77	0.78	0.77	0.76	0.83
Enrolment in education	0.11	0.77	0.81	0.78	0.73	0.86
GDP *per capita*	0.33	0.85	0.84	0.84	0.84	0.88

†Nominal weights w_i; main effects S_i, $S_{i,\text{lin}} := S_i(\infty)$ (linear fit), $S_{i,\text{CV}} := S_i(h_{\text{CV}})$, $S_{i,\text{DPI}} := S_i(h_{\text{DPI}})$, $S_{i,\min} := \min_{h\in\mathcal{H}} S_i(h)$ and $S_{i,\max} := \max_{h\in\mathcal{H}} S_i(h)$.

Table 3. Maximum discrepancy statistic d_m for various choices of the bandwidth h in the main effect estimator $S_i(h)$†

Index	$d_{m,\text{DPI}}$	$d_{m,\text{CV}}$	$d_{m,\text{lin}}$	$d_{m,\min}$	$d_{m,\max}$
ARWU	0.36	0.36	0.31	0.26	0.50
THES	0.41	0.42	0.34	0.29	0.55
2009 HDI	0.59	0.63	0.57	0.50	0.69
2010 HDI	0.06	0.07	0.09	0.03	0.13
IAG	0.29	0.34	0.42	0.13	0.57
SSI	0.85	0.91	0.95	0.38	0.98

†Minimum and maximum values were obtained by considering all possible combinations of S_{i,l_i}-values, where $l_i \equiv \min, \max$.

The hypothesis of linearity is not rejected for two indicators for the ARWU and for four indicators for the THES index, when evaluating the tests at h_{DPI} and h_{CV}. The two indicators for the ARWU are those with the highest proportion of values equal to 0, which were discarded in the choice of bandwidth; the numbers of valid cases n are 198 and 135 respectively. This may reflect the fact that it is more difficult to reject linearity with smaller samples. The indicators used in the THES ranking instead do not have so many 0 values; also here, however, we find that $E(y|x_i)$ is approximately linear for four indicators.

4.1.2. 2009 human development index

The HDI (see United Nations Development Programme (2009)) summarizes human devel-

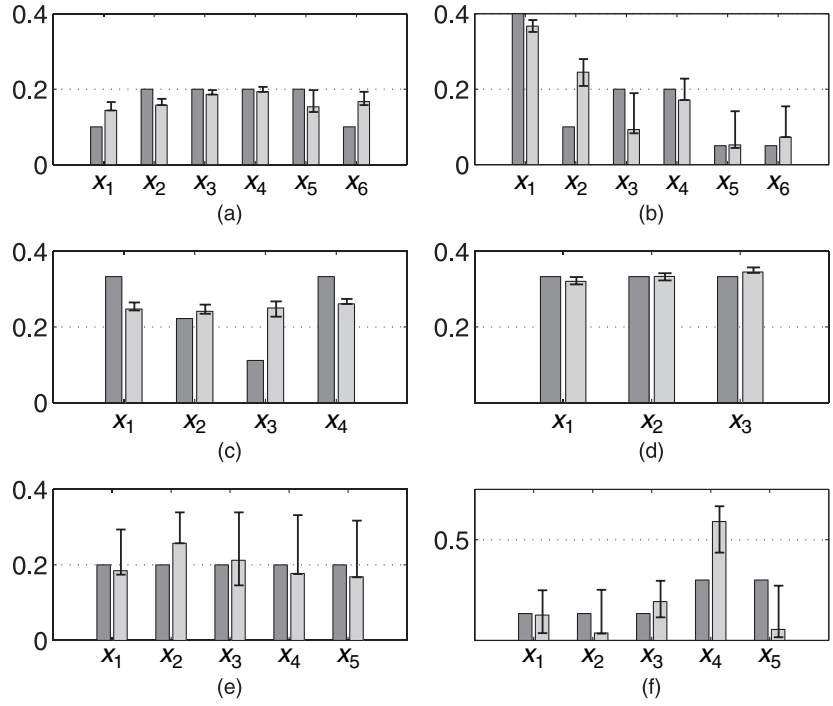

Fig. 5. Comparison of normalized main effects S_i^* (□) and w_i (■) ($S_i^* := S_i/c$ with $c := \Sigma_{t=1}^{k} S_t$ and $S_i := S_{i,\mathrm{CV}}$; the indicators x_i are numbered consecutively as in Tables 1, 2, 4 and 5; bounds for S_i^* are constructed as $S_{i,\min}/c$ and $S_{i,\max}/c$): (a) ARWU; (b) THES; (c) 2009 HDI; (d) 2010 HDI; (e) IAG; (f) SSI

opment in 182 countries on the basis of four indicators: a long healthy life measured by life expectancy at birth (weight 1/3), knowledge measured by adult literacy rate (weight 2/9) and combined primary, secondary and tertiary education gross enrolment ratio (weight 1/9) and a decent standard of living measured by the GDP *per capita* (weight 1/3). Raw data in the four indicators are normalized by using the min–max approach to be in [0, 1]. The 2009 HDI score is the weighted average of the four normalized indicators. Because data on the adult literacy rate were missing for several countries, we analysed data only for the countries without missing data; this gave a total of 142 countries. The four indicators present strong correlations that range from 0.70 to 0.81 and an average bivariate correlation of 0.74.

Nominal weights and estimates of the main effects are given in the bottom panel of Table 2, and the choice of bandwidth is given in Table 1. The maximum discrepancy is given in Table 3 and a graphical comparison of nominal weights and estimates of the main effects is provided in Fig. 5. Table 1 reports evidence on the choice of bandwidth h and the p-values for the linearity test, at the values h_{DPI} and h_{CV} of the smoothing parameter h.

Both the main effects S_i and the Pearson correlation coefficients reveal a relatively balanced effect of the four indicators life expectancy, GDP *per capita*, enrolment in education and adult literacy on the variance of the HDI scores, with the adult literacy indicator being slightly less important. It would seem that the HDI depends more equally on its four variables than the weights that were assigned by the developers would imply. For example, if one could fix adult literacy the variance of the HDI scores would on average be reduced by 77% (CV estimate), whereas by fixing the most influential indicator, GDP *per capita*, the variance reduction would be 84% on average.

Table 4. Main effects at pillar level†

	w_i	$S_{i,\mathrm{lin}}$	$S_{i,\mathrm{CV}}$	$S_{i,\mathrm{DPI}}$	$S_{i,\min}$	$S_{i,\max}$
2010 HDI						
Life expectancy	0.33	0.82	0.84	0.84	0.81	0.86
Education	0.33	0.86	0.87	0.86	0.84	0.89
GDP *per capita*	0.33	0.90	0.90	0.90	0.89	0.93
IAG						
Safety and security	0.20	0.52	0.54	0.63	0.51	0.87
Rule of law and corruption	0.20	0.77	0.76	0.78	0.76	1.00
Participation and human rights	0.20	0.44	0.63	0.68	0.43	1.00
Sustainable economic opportunity	0.20	0.52	0.52	0.56	0.52	0.98
Human development	0.20	0.50	0.50	0.55	0.49	0.94
SSI						
Personal development	0.13	0.05	0.14	0.17	0.04	0.27
Healthy environment	0.13	0.04	0.04	0.07	0.04	0.27
Well-balanced society	0.13	0.13	0.21	0.21	0.12	0.32
Sustainable use of resources	0.30	0.48	0.64	0.64	0.47	0.72
Sustainable world	0.30	0.02	0.06	0.10	0.02	0.29

†Nominal weights w_i; main effects S_i, $S_{i,\mathrm{lin}} := S_i(\infty)$ (linear fit), $S_{i,\mathrm{CV}} := S_i(h_{\mathrm{CV}})$, $S_{i,\mathrm{DPI}} := S_i(h_{\mathrm{DPI}})$, $S_{i,\min} := \min_{h\in\mathcal{H}} S_i(h)$, $S_{i,\max} := \max_{h\in\mathcal{H}} S_i(h)$.

One might suspect that it was precisely the developers' intention, when assigning nominal weights 11% and 33% to these two variables respectively, to make them equally important on the basis of the S_i-measure; however, this is not stated explicitly in the index documentation report United Nations Development Programme (2009). Overall, there is considerable discrepancy between the nominal weights that were assigned to the four indicators and their respective main effects in the 2009 HDI ($d_{m,\mathrm{CV}} = 0.63$).

The analysis of the 2009 HDI illustrates vividly that assigning unequal weights to the indicators is not a sufficient condition to ensure unequal importance. Although the 2009 HDI developers assigned weights varying between 11% and 33%, all four indicators are roughly equally important. The scatter plots in Figs 1–4 help to visualize the situation. In cases like this, where the variables are strongly and roughly equally correlated with the overall index, each of them ranks the countries roughly equally, and the weights are little more than cosmetic.

4.2. Importance at the pillar level

The issue of weighting is particularly fraught with normative implications in the case of pillars. As mentioned above, pillars in composite indicators are often given equal weights on the grounds that each pillar represents an important—possibly normative—dimension which could not and should not be seen to have more or less weight than the stipulated fraction. The discrepancy measure that is presented here can be of particular relevance and interest to gauge the quality of a composite indicator with respect to this important assumption. Here we consider the 2010 version of the HDI, the IAG and the SSI.

4.2.1. 2010 human development index

In this section we analyse the 2010 version of the HDI at pillar level, covering 169 countries. From the methodological viewpoint the main novelty in this version of the index is the use of a geometric—as opposed to an arithmetic—mean in the aggregation of the three pillars. The three

Table 5. Bandwidth choice at pillar level†

	h_{CV}	p_{CV}	h_{DPI}	p_{DPI}	n
2010 HDI					
Life expectancy	0.08	0.00	0.07	0.00	169
Education	0.02	0.09	0.06	0.21	169
GDP *per capita*	0.05	0.00	0.06	0.00	169
IAG					
Safety and security	17.69	0.15	3.31	0.45	53
Rule of law and corruption	25.05‡	0.30	4.75	0.94	53
Participation and human rights	4.89	0.08	2.85	0.41	53
Sustainable economic opportunity	22.14	0.09	4.21	0.51	53
Human development	25.05‡	0.17	3.42	0.87	53
SSI					
Personal development	0.69	0.00	0.37	0.00	151
Healthy environment	25.05‡	0.41	0.49	0.69	151
Well-balanced society	0.69	0.00	0.42	0.01	151
Sustainable use of resources	0.30	0.00	0.30	0.00	150
Sustainable world	0.86	0.00	0.38	0.01	151

†Bandwidth h, $h_{i,CV}$, $h_{i,DPI}$; p-values for the linearity test, $p_{i,CV}$ (p-value for $h_{i,CV}$), $p_{i,DPI}$ (p-value for $h_{i,DPI}$); n is the number of observations with $x_{ji} > 0$ used for CV and DPI.
‡Right-hand end of the grid $\mathcal{H}$.

pillars cover health (life expectancy at birth) x_{life}, education x_{edu} and income (gross national income *per capita*) x_{inc}. Education is the combination of two variables, namely mean years of schooling and expected years of schooling (United Nations Development Programme, 2010). The 2010 HDI index y is computed as

$$y = (x_{life} x_{edu} x_{inc})^{1/3}$$

where all three dimensions have equal weights. The reason for this change of aggregation scheme is to introduce an element of 'imperfect substitutability across all HDI dimensions', i.e. to reduce the compensatory nature of the linear aggregation; see United Nations Development Programme (2010), page 216.

Nominal weights and estimates of the main effects are given in the first panel of Table 4, whereas the choice of bandwidth is given in Table 5. The maximum discrepancy is given in the fourth row of Table 3 and a graphical comparison of nominal weights and estimates of the main effects is provided in Fig. 5.

Overall, the 2010 HDI shows very little discrepancy between the goals of equal importance of the three pillars and the main effects. In fact all three pillars have a similar effect on the index variance (roughly 84–90%). Hence, in this case the relative nominal weights are approximately equal to the relative effect of the pillars on the index variance. Such a correspondence is of value because it indicates that no pillar impacts too much or too little on the variance of the index as compared with its 'declared' equal importance. Compared with the other examples discussed, the 2010 HDI is the most consistent in this respect ($d_{m,CV} = 0.07$). The linearity tests reveal that the role of education is approximately linear within the index, despite the multiplicative aggregation scheme.

To assess the effect of the choice of the aggregation scheme on the index balance, we also perform a counterfactual analysis of the 2010 HDI using linear aggregation of the three dimensions. We find that this choice does not affect the relative importance of dimensions,

as these have comparable variances and covariances. Hence the 2010 HDI would have been balanced also under a linear aggregation scheme. This, however, does not detract from the conceptual appeal of imperfect substitutability that is implicit in geometric aggregation.

4.2.2. Index of African governance

The IAG was developed by the Harvard Kennedy School; see Rotberg and Gisselquist (2008); for a validation study see Saisana *et al.* (2009). In the 2008 version of the index, 48 African countries are ranked according to five pillars:

(a) safety and security,
(b) rule of law, transparency and corruption,
(c) participation and human rights,
(d) sustainable economic opportunity and
(e) human development.

The five pillars are described by 14 subpillars that are in turn composed of 57 indicators in total (in a mixture of qualitative and quantitative variables). Raw indicator data were normalized by using the min–max method on a scale from 0 to 100. The five pillar scores per country were calculated as the simple average of the normalized indicators. Finally, the IAG scores were calculated as the simple average of the five pillar scores. The five pillars have correlations that range from 0.096 to 0.76 and an average bivariate correlation of 0.45. Three pairwise correlations (involving participation and human rights and either sustainable economic opportunity or human development or safety and security) are not statistically significant at the 5% level.

Nominal weights and main effects are given in Table 4 and in Fig. 5, whereas the choice of bandwidth is reported in Table 5 and the discrepancy statistics in Table 3. The main conclusions are summarized as follows: the IAG is a good example of the situation that was discussed in Section 1 whereby all pillars represent important normative elements which by design should be equally important in the developers' intention. Overall the IAG appears to be balanced with respect to four pillars that have a similar effect on the index variance (roughly 50−63%), but the fifth pillar on the rule of law is more influential than conceptualized ($S_i = 76\%$; CV estimate). The IAG has a discrepancy statistic $d_{m,\mathrm{CV}} = 0.34$.

The linearity tests in Table 5 suggest that there is no statistical evidence against linearity for all the five indicators. Hence one could calculate S_i here as R_i^2.

4.2.3. Sustainable society index

The SSI has been developed by the Sustainable Society Foundation for 151 countries and it is based on a definition of sustainability of the Brundtland Commission (van de Kerk and Manuel, 2008). Also in this example, the five pillars of the index represent normative dimensions which are, however, considered of different importance: personal development (weight 1/7), healthy environment (1/7), well-balanced society (1/7), sustainable use of resources (2/7) and sustainable world (2/7). These five pillars are described by 22 indicators. Raw indicator data were normalized by using the min–max method on a scale from 0 to 10. The five pillars were calculated as the simple average of the normalized indicators. The SSI scores were calculated as the weighted average of the five pillar scores.

One can note that the linearity test suggests that for the second pillar 'healthy environment' there is no evidence against linearity of its relationship to the SSI. The five pillars have correlations that range from −0.62 to 0.75, where negative correlations between pillars are

generally undesired, as they suggest the presence of trade-offs between pillars (for example economic performance can only come with an environmental cost). Such trade-offs within index dimensions are a reminder of the danger of compensability between dimensions.

For the SSI, there are notable differences between declared and variance-based importance for the five pillars. The different association between a pillar and the overall index can also be grasped visually in Fig. 5. The two pillars on 'sustainable use of resources' and on 'sustainable world' are meant to be equally important according to the nominal weights (2/7 each), whereas the main effects suggest that the variance reduction that is obtained by fixing the former is 67% compared with merely 9% by fixing the latter. This strong discrepancy is due to the significant negative correlations between the SSI pillars. Overall, the level of maximal discrepancy of the SSI is the highest of the examples discussed ($d_{m,\mathrm{CV}} = 0.91$). The authors and the developers of the SSI have been communicating on this issue, and the 2010 version of the SSI index appears considerably improved; see `http://www.ssfindex.com/ssi/`.

4.3. Reverse engineering the weights

Applying the reverse engineering exercise that is described in Section 3.4 and Appendix A to our test cases (except for the case of the 2010 HDI that has low maximal discrepancy between relative weights and relative importance for the three pillars, and it is not obtained by the linear aggregation scheme (1)), we find that to achieve a relative effect of the indicators (or pillars) (as measured by the square of the Pearson correlation coefficient R_i^2) that equals the relative 'declared' importance of the indicators, negative nominal weights are involved in all studies except for the SSI. In the case of the SSI, to guarantee that the two pillars on sustainable use of resources and sustainable world are twice as important as the other three pillars, the nominal weights to be assigned to them are 0.19 for personal development, 0.16 for healthy environment, 0.07 for well-balanced society, 0.16 for sustainable use of resources and 0.41 for sustainable world. For all other cases, the data correlation structure does not allow the developers to achieve the stated relative importance by choosing positive weights.

5. Conclusions

According to many—including some of the authors of the Stiglitz report (see Stiglitz *et al.* (2009))—composite indicators have serious shortcomings. The debate among those who prize their pragmatic nature in relation to pragmatic problems (see Hand (2009)) and those who consider them an aberration is unlikely to be settled soon; see Saltelli (2007) for a review of pros and cons. Still these measures are pervasive in the public discourse and represent perhaps the best-known face of statistics in the eyes of the general public and media.

One might muse that what official statistics are to the consolidation of the modern nation state (see Hacking (1990)) composite indicators are to the emergence of post modernity—meaning by this the philosophical critique of the exact science and rational knowledge programme of Descartes and Galileo; see Toulmin (1990), pages 11–12. On a practical level, it is undeniable that composite indicators give voice to a plurality of different actors and normative views. Stiglitz *et al.* (2009) remarked (page 65) that

> 'The second [argument against composite indicators] is a general criticism that is frequently addressed at composite indicators, i.e. the arbitrary character of the procedures used to weight their various components. ... The problem is not that these weighting procedures are hidden, non-transparent or non-replicable—they are often very explicitly presented by the authors of the indices, and this is one of the strengths of this literature. The problem is rather that their normative implications are seldom made explicit or justified.'

The analysis of this paper shows that, although the weighting procedures are often very explicitly presented by the authors of the indices, the implications of these are neither fully understood nor assessed in relation to the normative implications. This paper proposes a variance-based tool to measure the internal discrepancy of a composite indicator between target and effective importance.

Our main conclusions can be summarized as follows. For transparency and simplicity, composite indicators are most often built by using linear aggregation procedures which are fraught with the difficulties that were described in Section 1: practitioners know that weights cannot be used as importance, although they are precisely elicited as if they were. Weights are instead measures of substitutability in linear aggregation. The error is particularly severe when a variable's weight substantially deviates from its relative strength in determining the ordering of the units (e.g. countries) being measured.

Pearson's correlation ratio (or main effect) that is suggested in this paper is a suitable measure of importance of a variable (be it indicator or pillar) because

(a) it offers a precise definition of importance (i.e. 'the expected reduction in variance of the composite indicator that would be obtained if a variable could be fixed'),
(b) it can be used regardless of the degree of correlation between variables,
(c) it is model free, in that it can be applied also in non-linear aggregations, and finally
(d) it is not invasive, in that no changes are made to the composite indicator or to the correlation structure of the indicators.

Because of property (a) and the fact that it takes the whole covariance structure into account, the main effect can also be useful to prioritize variables on which a country or university, or whatever units are being rated, could intervene to improve its overall score. Note that the indicator with highest main effect is not necessarily the indicator in which the country scores the worst.

The main effects approach can complement the techniques for robustness analysis applied to composite indicators thus far seen in the literature; see for example Saisana *et al.* (2005, 2011) and Organisation for Economic Co-operation and Development (2008). The approach that is described in this paper does not need an explicit modelling of error propagation but it is simply based on the data as produced by developers.

The discrepancy statistic based on the absolute error between ratios of the main effects and of the corresponding target relative importance provides a pragmatic answer to the research question that is posed in this paper. Relative main effects are variance based, and hence they are ratios of quadratic forms of nominal weights, whereas target relative importance is often deduced as ratios of nominal weights. Comparing them via the discrepancy statistic is a way of comparing these two importance measures, one of which is stated *ex ante* as a target and the other that is computed *ex post*; this allows us to see how close the two measures are in practice.

The discrepancy statistic has been effective in the six examples that were discussed, in that it allowed an analytic judgement about the discrepancy in the assignment of the weights in two well-known measures of higher education performance ($d_m = 0.42$ for the THES index *versus* $d_m = 0.36$ for the ARWU), two versions of an HDI ($d_m = 0.63$ for the 2009 HDI and $d_m = 0.02$ for the 2010 HDI), one index of governance ($d_m = 0.34$ for the IAG) and one index of sustainability ($d_m = 0.86$ for the SSI).

Our reverse engineering analysis shows that in most cases it is not possible to find nominal weights that would give the desired importance to variables. This can be a useful piece of information to developers and might induce a deeper reflection on the cost of the simplification that is achieved with linear aggregation. Developers could thus

(a) avoid associating nominal weights with importance but inform users of the relative importance of the variables or pillars, using statistics such as those presented in this paper,
(b) abstain from aggregating pillars when these display important trade-offs which make it difficult to give them target weights in an aggregated index,
(c) reconsider the aggregation scheme, moving from the linear scheme (which is fully compensatory) to a partially or fully non-compensatory alternative, such as a Condorcet like (or approximate Condorcet) approach, where weights would fully play their role as a measure of importance (see Munda (2008)) and
(d) assess different weighting strategies, to select the strategy that leads to a minimum discrepancy statistic between target weights and variables' importance.

Acknowledgements

We thank, without implicating, Beatrice d'Hombres, Giuseppe Munda, the Associate Editor and two referees for useful comments. The views expressed are those of the authors and not of the European Commission or the University of Insubria.

Appendix A: Solution to the inverse problem

In the linear case, the ratio S_i/S_1 equals the ratio of squares of Pearson's correlation coefficients R_i^2/R_1^2; this is a function $H_i(\mathbf{w})$ of $\mathbf{w} := (w_1, \ldots, w_k)$ and of the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{x} := (x_1, \ldots, x_k)'$. We find

$$H(\mathbf{w}) = \frac{(\mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{w})^2\sigma_{11}}{(\mathbf{e}_1'\boldsymbol{\Sigma}\mathbf{w})^2\sigma_{ii}},$$

where $\mathbf{e}_i$ is the ith column of the identity matrix of order k and σ_{ii} is the ith variance on the diagonal of $\boldsymbol{\Sigma}$. We wish to make $H_i(\mathbf{w})$ equal to a preselected value z_i^2 for all i,

$$H_i(\mathbf{w}) = z_i^2, \qquad i = 1, \ldots, k, \tag{8}$$

and seek to find a solution $\mathbf{w} \in \mathbb{R}^k$ to this problem such that the nominal weights sum to 1, i.e.

$$\mathbf{1}'\mathbf{w} = 1. \tag{9}$$

We show that this solution is unique and it is given by equation (7) in the text, where $\mathbf{g}$ is a vector with ith entry equal to $g_i := z_i\sqrt{(\sigma_{ii}/\sigma_{11})} > 0$ and $\mathbf{1}$ is a k-vector of 1s.

By construction $g_1 = 1$. We have that expression (8) can be written as

$$\mathbf{e}_1'\boldsymbol{\Sigma}\mathbf{w} - \frac{1}{g_i}\mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{w} = \mathbf{0},$$

or, setting $\mathbf{G} := \mathrm{diag}(1, 1/g_2, \ldots, 1/g_k)$ and $\mathbf{F} := \mathbf{1}\mathbf{e}_1'$, as $(\mathbf{F} - \mathbf{G})\boldsymbol{\Sigma}\mathbf{w} = \mathbf{0}$. This shows that $\boldsymbol{\Sigma}\mathbf{w}$ should be selected in the right null space of $\mathbf{F} - \mathbf{G}$. We observe that

$$\mathbf{F} - \mathbf{G} = \begin{pmatrix} 0 & 0 & & 0 \\ 1 & -1/g_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 1 & 0 & & -1/g_k \end{pmatrix}$$

whose right null space $\mathcal{A}$ is one dimensional; moreover $\mathcal{A}$ is spanned by $\mathbf{g} := (1, g_2, \ldots, g_k)'$. Hence $\boldsymbol{\Sigma}\mathbf{w} = \mathbf{g}c$ for a non-zero c or $\mathbf{w} = \boldsymbol{\Sigma}^{-1}\mathbf{g}c$. Substituting this expression in equation (9), we find that $1 = \mathbf{1}'\mathbf{w} = \mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{g}c$, which implies that $c = 1/\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{g}$. We hence conclude that the weights that satisfy equation (8) are given by equation (7), and that they are unique.

References

Agrast, M. D., Botero, J. C. and Ponce, A. (2010) Rule of law index 2010. *Technical Report*. World Justice Project, Washington DC. (Available from `http://worldjusticeproject.org/`.)

Balinski, M. and Laraki, R. (2010) *Majority Judgment: Measuring, Ranking and Electing*. Cambridge: MIT Press.

Bandura, R. (2008) A survey of composite indices measuring country performance: 2008 update. *Technical Report*. United Nations Development Programme, Office of Development Studies, New York.

Billaut, J. C., Bouyssou, D. and Vincke, P. (2010) Should you believe in the Shanghai ranking? *Scientometrics*, **84**, 237–263.

Bird, S. M., Cox, D., Farewell, V. T., Goldstein, H., Holt, T. and Smith, P. C. (2005) Performance indicators: good, bad, and ugly. *J. R. Statist. Soc.* A, **168**, 1–27.

Bowman, A. W. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis: a Kernel Approach with S-Plus Illustrations*. Oxford: Clarendon.

Boyssou, D., Marchant, T., Pirlot, M. and Tsoukiàs, A. (2006) *Evaluation and Decision Models with Multiple Criteria: Stepping Stones for the Analyst*. Berlin: Springer.

Center for World-Class Universities (2008) Academic ranking of world universities—2008. *Technical Report*. Center for World-Class Universities, Institute of Higher Education, Shanghai Jiao Tong University, Shanghai. (Available from `http://www.arwu.org/`.)

Decancq, K. and Lugo, M. (2010) Weights in multidimensional indices of well-being: an overview. *Econmetr. Rev.*, to be published, doi 10.1080/07474938.2012.69041.

Freedom House (2011) Freedom of the press 2011. *Technical Report*. Freedom House, Washington DC. (Available from `http://www.freedomhouse.org`.)

Hacking, I. (1990) *The Taming of Chance*. Cambridge: Cambridge University Press.

Hand, D. (2009) *Measurement Theory and Practice: the World through Quantification*. Chichester: Wiley.

Hendrik, W., Howard, C. and Maximilian, A. (2008) Consequences of data error in aggregate indicators: evidence from the human development index. *Technical Report*. Department of Agricultural and Resource Economics, University of California at Berkeley, Berkeley.

van de Kerk, G. and Manuel, A. R. (2008) A comprehensive index for a sustainable society: the SSI, Sustainable Society Index. *J. Ecol. Econ.*, **66**, 228–242.

Leckie, G. and Goldstein, H. (2009) The limitations of using school league tables to inform school choice. *J. R. Statist. Soc.* A, **172**, 835–851.

Li, G., Rabitz, H., Yelvington, P., Oluwole, O., Bacon, F., Kolb, C. and Schoendorf, J. (2010) Global sensitivity analysis for systems with independent and/or correlated inputs. *J. Phys. Chem.*, **114**, 6022–6032.

Munda, G. (2008) *Social Multi-criteria Evaluation for a Sustainable Economy*. Berlin: Springer.

Munda, G. and Nardo, M. (2009) Non-compensatory/non-linear composite indicators for ranking countries: a defensible setting. *Appl. Econ.*, **41**, 1513–1523.

Nardo, M. (2009) Product market regulation: robustness and critical assessment 1998-2003-2007—how much confidence can we have on PMR ranking? *Technical Report EUR 23667*. Joint Research Centre, European Commission, Ispra.

Nicoletti, G., Scarpetta, S. and Boylaud, O. (2000) Summary indicators of product market regulation with an extension to employment protection legislation. *Working Paper 226*. Economics Department, Organisation for Economic Co-operation and Development, Paris. (Available from `http://dx.doi.org/10.1787/215182844604`.)

Organisation for Economic Co-operation and Development (2008) *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Paris: Organisation for Economic Co-operation and Development.

Pearson, K. (1905) *Mathematical Contributions to the Theory of Evolution, Drapers' Company Research Memoirs*, vol. XIV, *On the General Theory of Skew Correlation and Non-linear Regression*. London: Dulau.

Plischke, E. (2010) An effective algorithm for computing global sensitivity indices (EASI). *Reliab. Engng Syst. Safty*, **95**, 354–360.

Ratto, M. and Pagano, A. (2010) Recursive algorithms for efficient identification of smoothing spline anova models. *Adv. Statist. Anal.*, **94**, 367–388.

Ratto, M., Pagano, A. and Young, P. (2007) State dependent parameter metamodelling and sensitivity analysis. *Comput. Phys. Communs*, **177**, 863–876.

Ravallion, M. (2010) Troubling tradeoffs in the human development index. *Policy Research Working Paper 5484*. Development Research Group, World Bank, Washington DC.

Reporters Sans Frontières (2011) Press freedom index. *Technical Report*. Reporters Without Borders, Paris. (Available from `http://en.rsf.org/press-freedom-index-2010,1034.html`.)

Rotberg, R. and Gisselquist, R. (2008) Strengthening African governance: Ibrahim index of African governance; results and rankings 2008. *Technical Report*. Kennedy School of Government, Harvard University, Boston.

Ruppert, A., Sheather, S. J. and Wand, M. P. (1995) An effective bandwidth selector for local least squares regression. *J. Am. Statist. Ass.*, **90**, 1257–1270.

Ruppert, D. and Wand, M. P. (1994) Multivariate locally weighted least squares regression. *Ann. Statist.*, **22**, 1346–1370.

Saaty, T. (1980) *The Analytic Hierarchy Process*. New York: McGraw-Hill.

Saaty, T. L. (1987) The analytic hierarchy process: what it is and how it is used. *Math. Modllng*, **9**, 161–176.
Saisana, M., Annoni, P. and Nardo, M. (2009) A robust model to measure governance in African countries. *Technical Report EUR 23773*. Joint Research Centre, European Commission, Ispra.
Saisana, M., d'Hombres, B. and Saltelli, A. (2011) Rickety numbers: volatility of university rankings and policy implications. *Res. Poly*, **40**, 165–177.
Saisana, M., Saltelli, A. and Tarantola, S. (2005) Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *J. R. Statist. Soc.* A, **168**, 307–323.
Saltelli, A. (2002) Making best use of model valuations to compute sensitivity indices. *Comput. Phys. Communs*, **145**, 280–297.
Saltelli, A. (2007) Composite Indicators between analysis and advocacy. *Socl Indic. Res.*, **81**, 65–77.
Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. and Tarantola, S. (2010) Variance based sensitivity analysis of model output: design and estimator for the total sensitivity index. *Comput. Phys. Communs*, **181**, 259–270.
Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008) *Global Sensitivity Analysis—the Primer*. Chichester: Wiley.
Saltelli, A. and Tarantola, S. (2002) On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. *J. Am. Statist. Ass.*, **97**, 702–709.
Sobol', I. (1993) Sensitivity analysis for non-linear mathematical models. *Math. Modlng Computnl Expt*, **1**, 407–414 (Engl. transl.).
Stanley, J. and Wang, M. (1968) Differential weighting: a survey of methods and empirical studies. *Technical Report*. Johns Hopkins University, Baltimore.
Stiglitz, J. E., Sen, A. and Fitoussi, J. (2009) Report by the Commission on the Measurement of Economic Performance and Social Progress. *Technical Report*. Commission on the Measurement of Economic Performance and Social Progress, Paris. (Available from `www.stiglitz-sen-fitoussi.fr`.)
Tarantola, S., Gatelli, D. and Mara, T. (2006) Random balance designs for the estimation of first order global sensitivity indices. *Reliab. Engng Syst. Safty*, **91**, 717–727.
Times Higher Education Supplement (2008) World university rankings. *Technical Report*. Times Higher Education Supplement, London. (Available from `http://www.timeshighereducation.co.uk`.)
Toulmin, S. (1990) *Cosmopolis—the Hidden Agenda of Modernity*. Chicago: University of Chicago Press.
United Nations Development Programme (2009) Human development report 2009. *Technical Report*. United Nations Development Programme, New York. (Available from `http://hdr.undp.org/en/reports/`.)
United Nations Development Programme (2010) Human development report 2010. *Technical Report*. United Nations Development Programme, New York. (Available from `http://hdr.undp.org/en/reports/`.)
Wang, M. W. and Stanley, J. C. (1970) Differential weighting: a review of methods and empirical studies. *Rev. Educ. Res.*, **40**, 663–705.
Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd edn. Cambridge: MIT Press.
World Economic Forum (2010) The Global Competitiveness Report 2010-2011. *Technical Report*. World Economic Forum, Geneva. (Available from `http://www.weforum.org/`.)
Xu, C. and Gertner, G. (2011) Understanding and comparisons of different sampling approaches for the fourier amplitudes sensitivity test (fast). *Computnl Statist. Data Anal.*, **55**, 184–198.