MEASURING THE IMPORTANCE OF VARIABLES IN COMPOSITE INDICATORS

William Becker[1], Michaela Saisana[1], Paolo Paruolo[1], Andrea Saltelli[2]

[1] European Commission – Joint Research Centre, Econometrics and Applied Statistics Unit, COIN Group (e-mail: william.becker@jrc.it, michaela.saisana@jrc.it, paolo.paruolo@jrc.it)

[2] Centre for the Study of the Sciences and the Humanities - University of Bergen (e-mail: andrea.saltelli@svt.uib.no)

KEYWORDS: Composite indicators, nonlinear regression, correlation

1 Introduction

Composite indicators are aggregations of measurable variables (indicators) that aim to quantify underlying concepts that are not directly observable, such as competitiveness, freedom of press or climate hazards. Composite indicators, otherwise referred to as performance indices, are employed for many purposes, including policy monitoring.

Sensitivity analysis can be 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.

This work highlights how the relative importance of input variables should not be confused with the nominal weights, or transformations of variables, that are sometimes used when constructing composite indicators. In fact the importance and the relative weights rarely coincide due to correlations between input variables. This is demonstrated empirically in this work, and is further explained in Paruolo *et al.*, 2013. Here, we use nonlinear regression to estimate the Pearson correlation ratio as a measure of importance. The methodology presented here is used to investigate the weights of three composite indicators.

2 Measures of Importance

Consider the case of a composite indicator y calculated using some aggregation rule over k variables $\{x_i\}_{i=1}^{k}$. An example could be a weighted arithmetic

average, i.e.

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

where x_{ji} is the normalised score of individual j (e.g., country) and w_i is the nominal weight assigned to variable x_i.

In order to estimate the true influence of each input on the composite indicator, the proposal is to use the *correlation ratio*, or *first order sensitivity index*, S_i, $i = 1, 2, ..., k$. This measure has the added value over linear measures of dependence, such as R_i^2, that it can capture nonlinear dependence. It can be interpreted as the expected variance reduction of the composite indicator, if a given variable were fixed. The correlation ratio, traditionally denoted as η_i^2 by Pearson, 1905, is defined as:

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

where $\mathbf{x}_{\sim i}$ is defined as the vector containing all the variables $(x_1, \ldots, x_k)$ except variable x_i and $\mathrm{E}_{\mathbf{x}_{\sim i}}(y \mid x_i)$ denotes the conditional expectation of y given x_i.

The conditional expectation $\mathrm{E}(y \mid x_i)$ is known as the *main effect* of x_i on y, and is a function of x_i which can be estimated by performing a (nonlinear) regression of y on x_i, which can then be used to estimate S_i.

There are many methods available to estimate $\mathrm{E}(y \mid x_i)$. In this work, we use two nonlinear regression approaches: penalised splines and local-linear regression.

Penalised splines are a form of linear (linear in the parameters) regression, based on a weighted combination of polynomial functions controlled by a tuning parameter, which effectively controls the smoothness of the spline fit through the data. Local-linear regression similarly consists of averaging a number of linear regressions, centred at different values of x_i, but weighted by Gaussian kernel functions.

In both approaches there is a smoothing parameter which is found by cross-validation. In many situations, penalised splines and local-linear regression will produce very similar fits to the data, although in the presence of strongly nonlinear and/or heteroscedastic data (which is not uncommon in composite indicators) they may also differ substantially. Nevertheless there is no obvious reason to choose one over the other, although splines tend to be faster to fit; this is only a significant advantage however when performing multiple regressions on large datasets.

3 Results and Conclusion

	x_i	w_i	R_i	R_i^2	$S_{i,\text{spl}}$	$S_{i,\text{ll}}$
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. $R_i = \text{corr}(x_{i,}, y)$*: correlation;* $S_{i,spl}$*: correlation ratio, spline;* $S_{i,ll}$*: correlation ratio, kernel.*

The three composite indicators investigated in this work were as follows: the Resource Governance Index (RGI), which aims to measure transparency and accountability in the oil, gas and mining sectors; the Financial Secrecy Index, which measures secrecy 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 common good of humanity. Due to space limitations, only the results for the RGI will be presented in this abstract.

Figure 1 shows the nonlinear regression fits to the Resource Governance Index, compared with linear regression. One can see that in the first three variables the nonlinear regression is quite similar to the linear fit, but in the fourth variable ("enabling environment") the fits differ markedly, with the spline giving a smooth fit and the local linear (kernel) regression giving a slightly rougher fit.

Table 1 shows the estimates of S_i, also compared to R_i^2, which demonstrate 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. The association between reporting practices and safeguards and quality control is very high (0.82, the highest among the components). If one could fix reporting practices, the variance of the RGI scores across the 58 countries would on average be reduced by 94% (kernel estimate). It is worth noting that despite the equal weights assigned to the other three components, their impact to the RGI variation differs: by fixing any of the other components, the variance reduction would be 83% for Safeguards and quality control, 70% for Enabling Environment and 67% for Institutional and Legal Setting.

A similar analysis was applied to the other two composite indicators described at the beginning of this section; the main findings are that in the Fi-

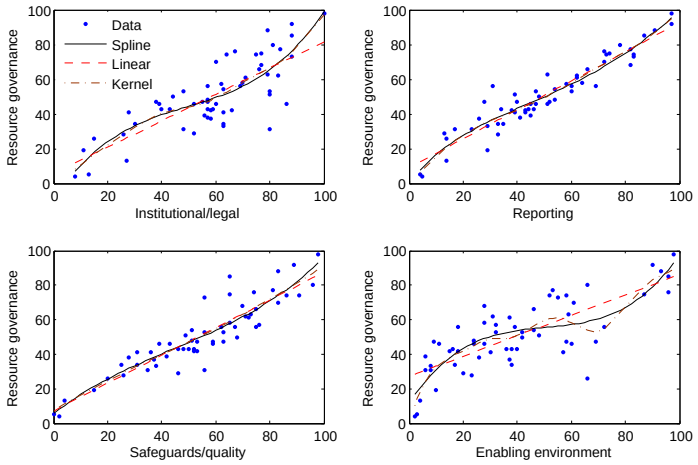

Figure 1. *Penalised spline, kernel and linear fits to Resource Governance Index.*

nancial Secrecy Index the influence of each variable is rather unbalanced due to very strong negative correlations, and similarly in the Good Country Index there are inputs which are far less influential than the developers' intentions. We suggest that the indicators could be more effectively weighted by finding weights which correspond to the desired correlation ratios, using for example an optimisation procedure.

Overall the nonlinear regression approaches here offer an added value to linear dependency measures, allowing a deeper insight into the influence of the inputs to composite indicators, usually finding that the intentions of the developers do not correspond with the reality.

References

PARUOLO, PAOLO, SAISANA, MICHAELA, & SALTELLI, ANDREA. 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.