Institute for the Protection and Security of the Citizen
Econometrics and Statistical Support to Antifraud Unit
I-21020 Ispra (VA) Italy

Tools for Composite Indicators Building

Prepared by

Michela Nardo, Michaela Saisana, Andrea Saltelli & Stefano Tarantola

(Applied Statistics Group)

2005 **EUR 21682 EN**

LEGAL NOTICE

The views expressed in this report are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information

A great deal of information on the European Union is available on the Internet. It can be accessed through the Europa server (http://europa.eu.int).

The Report is available online at http://farmweb.jrc.cec.eu.int/ci/bibliography.htm

EUR 21682 EN

Table of Contents

Foreword

Our society is changing so fast we need to know as soon as possible when things go wrong (Euroabstracts, 2003). This is where composite indicators enter into the discussion. A composite indicator is an aggregated index comprising individual indicators and weights that commonly represent the relative importance of each indicator. However, the construction of a composite indicator is not straightforward and the methodological challenges raise a series of technical issues that, if not addressed adequately, can lead to composite indicators being misinterpreted or manipulated. Therefore, careful attention needs to be given to their construction and subsequent use.

This document reviews the steps involved in a composite indicator's construction process and discusses the common pitfalls to be avoided. We stress the need for multivariate analysis prior to the aggregation of the individual indicators. We deal with the problem of missing data and with the techniques used to bring into a common unit the indicators that are of very different nature. We explore different methodologies for weighting and aggregating indicators into a composite and test the robustness of the composite using uncertainty and sensitivity analysis. Finally we show how the same information that is communicated by the composite indicator can be presented in very different ways and how this can influence the policy message.

Important note

The material presented here will eventually feed in a joint OECD-JRC Handbook of composite indicators building, expected to appear in fall 2005.

1. Introduction

Composite indicators are increasingly recognized as a useful tool for policy making and public communications in conveying information on countries' performance in fields such as environment, economy, society, or technological development. Composite indicators are much easier to interpret than trying to find a common trend in many separate indicators. Composite indicators have proven to be useful in ranking countries in benchmarking exercises. However, composite indicators can send misleading or non-robust policy messages if they are poorly constructed or misinterpreted. Andrew Sharpe (2004) notes:

"The aggregators believe there are two major reasons that there is value in combining indicators in some manner to produce a bottom line. They believe that such a summary statistic can indeed capture reality and is meaningful, and that stressing the bottom line is extremely useful in garnering media interest and hence the attention of policy makers. The second school, the non-aggregators, believe one should stop once an appropriate set of indicators has been created and not go the further step of producing a composite index. Their key objection to aggregation is what they see as the arbitrary nature of the weighting process by which the variables are combined."

In Saisana et al. (2005) one reads:

"[…] it is hard to imagine that debate on the use of composite indicators will ever be settled […] official statisticians may tend to resent composite indicators, whereby a lot of work in data collection and editing is "wasted" or "hidden" behind a single number of dubious significance. On the other hand, the temptation of stakeholders and practitioners to summarise complex and sometime elusive processes (e.g. sustainability, single market policy, etc.) into a single figure to benchmark country performance for policy consumption seems likewise irresistible."

Synthetically the main pros and cons of using composite indicators could be summarized as follows:

Pros of composite indicators
+ Summarise complex or multi-dimensional issues, in view of supporting decision-makers.
+ Are easier to interpret than trying to find a trend in many separate indicators.
+ Facilitate the task of ranking countries on complex issues in a benchmarking exercise.
+ Assess progress of countries over time on complex issues.
+ Reduce the size of a set of indicators or include more information within the existing size limit.
+ Place issues of countries performance and progress at the centre of the policy arena.
+ Facilitate communication with ordinary citizens and promote accountability.

Cons of composite indicators
- May send misleading policy messages, if they are poorly constructed or misinterpreted.
- May invite drawing simplistic policy conclusions, if not used in combination with the indicators.
- May lend themselves to instrumental use (e.g be built to support the desired policy), if the various stages (e.g. selection of indicators, choice of model, weights) are not transparent and based on sound statistical or conceptual principles.
- The selection of indicators and weights could be the target of political challenge.
- May disguise serious failings in some dimensions of the phenomenon, and thus increase the difficulty in identifying the proper remedial action.
- May lead wrong policies, if dimensions of performance that are difficult to measure are ignored.

A composite indicator is the mathematical combination of individual indicators that represent different dimensions of a concept whose description is the objective of the analysis (see Saisana and Tarantola, 2002). The construction of composite indicators involves stages where subjective judgement has to be made: the selection of indicators, the treatment of missing values, the choice of aggregation model, the weights of the indicators, etc. These subjective choices can be used to manipulate the results. It is, thus, important to identify the sources of subjective or imprecise assessment and use uncertainty and sensitivity analysis to gain useful insights during the process of composite indicators building, including a contribution to the indicators' quality definition and an appraisal of the reliability of countries' ranking.

We would point that composite indicators should never be seen as a goal per se. They should be seen, instead, as a starting point for initiating discussion and attracting public interest and concern. The aim of the present document is to provide guidance on how to ascertain that the process leading to the construction of a composite indicator meets certain quality objectives. The structure of this document is as follows: Section 2 describes the main issues related with the construction of composite indicators, which are then treated in detail in the following sections. Sections 3 to 5 deal with the statistical treatment of the set of indicators: multivariate analysis, imputation of missing data and normalization techniques aim at supplying a sound and defensible dataset. Section 6 gives the developers and users of composite indicators an introduction to the main weighting and aggregation procedures. Section 7 explores the merits of applying uncertainty and sensitivity analysis to increase transparency and make policy inference more defensible. Section 8 shows how different visualization strategies of the same composite indicator can convey different policy messages. The Technology Achievement Index (TAI), a composite indicator developed by the United Nations (Human Development Report, UN 2001), has been chosen as example to elucidate the various steps in the construction of a composite indicator and guide the reader into the different problems that may arise (a detailed description of the composite indicator is given in the Appendix).

2. Construction of composite indicators

The composite indicators' controversy can perhaps be put into context if one considers that indicators, and a fortiori composite indicators, are models, in the mathematical sense of the term. Models are inspired from systems (natural, biological, social) that one wishes to understand. Models are themselves systems, formal system at that. The biologist Robert Rosen (1991, Figure 2.1) noted that while a causality entailment structure defines the natural system, and a formal causality system entails the formal system, no rule of encoding the formal system given the real system, i.e. to move from perceived reality to model, was ever agreed.

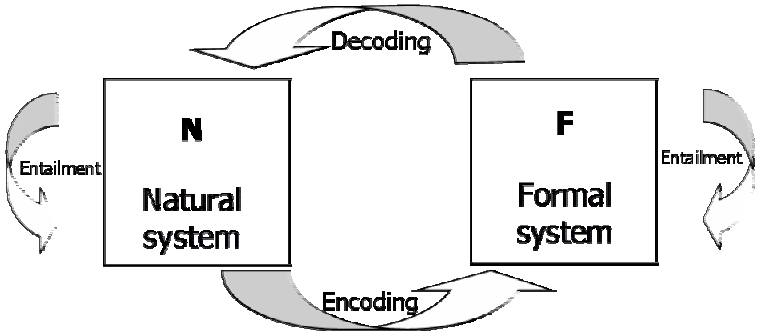


Figure 2.1, From Rosen 1991.

The formalization of the system generates an image, the **theoretical framework**, that is valid only within a given information space. As result, the model of the system will reflect not only (some of) the characteristics of the real system but also the choices made by the scientists on how to observe the reality. When building a model to describe a real-world phenomenon, formal coherence is a necessary property, yet not sufficient. The model in fact should fit objectives and intentions of the user, i.e. it must be the most appropriate tool for expressing the set of objectives that motivated the whole exercise. The choice of which sub-indicators to use, how those are divided into classes, whether a normalization method has to be used (and which one), the choice of the weighting method, and how information is aggregated, all these features stem from a certain perspective on the issue to be modelled. Reflexivity is thus an essential feature of a model since "*the observer and the observation are not separated [...] the way human kind approaches the problem is part of the problem itself.*" (Gough at al. 1998).

No matter how subjective and imprecise the theoretical framework is, it implies the recognition of the multidimensional nature of the phenomenon to be measured and the effort of specifying the single aspects and their interrelation. Most of the issues described with a composite indicator are complex problems, think to concepts like welfare, quality of education, or sustainability. Complexity is reflected by the multi-dimensionality and multi-scale representation of the issue.

The European Commission, for example, recognises the multi-dimensionality in the definition of sustainability claming that the social, environmental and economic dimensions must be dealt with together (European Commission, 'A Sustainable Europe for a Better World: a European Union Strategy for Sustainable Development' COM(2001)264 final of 15.05.2001). Defining sustainability within a multi-dimensional framework entails merging multidisciplinary point of views, all equally legitimate opinions of what is sustainability and how should be measured. Then, for each discipline, e.g. economics, sustainability can be measured at different (hierarchical) levels like economic agents, households, economic sectors, nations, European Union, or the entire planet. Synergies and conflicts, that would appear when sustainability is measured on a national or on a wider scale (think to policies related to the climate change), are likely to disappear at the local level where other aspects prevail. The change in scale might also produce contradictory implications and remedies all equally justifiable (e.g. windmills are desirable sources of clean energy at a national level but might produce social disputes in the local communities where windmills have to be placed).

Giampietro et al. (2004) notice that in complex issues the 'quality' of the theoretical framework depends on " three crucial challenges for the scientific community":

1. check the feasibility of the effect of the proposed [*framework*] in relation to different dimensions (technical, economic, social, political, cultural) and different scales: local (e.g. technical coefficients), medium (e.g. aggregate characteristics of large units) and large scales (e.g. trend analysis and benchmarks to compare trajectories of development)…. (italics added)
2. address several legitimate (and often contrasting) perspectives found among stakeholders on how to structure the problem….
3. handle in a credible way the unavoidable degree of uncertainty, or even worst, genuine ignorance associated to any multi-scale, multi-dimensional analysis of complex adaptive systems."

If we accept a definition of the theoretical framework requiring the integration of a broad set of (probably conflicting) points of view and the use of non-equivalent representative tools then the problem becomes to reduce the complexity in a measurable form. In other terms non-measurable issues like sustainability need to be replaced by intermediate objectives whose achievement can be observed and measured. The reduction into parts has limits when crucial properties of the entire system are lost: often the individual pieces of a puzzle hide the whole picture.

As suggested by Box (1979): *'all models are wrong, some are useful'*. The quality of a composite indicator is thus in its fitness or function to purpose. This is recognised by A. K. Sen (1989), Nobel prize winner in 1998, who was initially opposed to composite indicators but was eventually seduced by their ability to put into practice his concept of 'Capabilities' (the range of things that a person could do and be in her life) in the UN Human Development Index[1].

Although we cannot tackle here the vast issue of quality of statistical information, there is one aspect of the quality of composite indicators which we find essential for their use. This is the existence of a community of peers (be these individuals, regions, countries, facilities of various nature) willing to accept the composite indicators as their common yardstick based on their understanding of the issue. In discussing pedigree matrices for statistical information (see Section 2.2) Funtowicz and Ravetz note (in Uncertainty and Quality in Science for Policy, 1990)

"[…] any competent statistician knows that "just collecting numbers" leads to nonsense. The whole Pedigree matrix is conditioned by the principle that statistical work is (unlike some traditional lab research) a highly articulated social activity. So in "Definition and Standards" we put "negotiation" as superior to "science", since those on the job will know special features and problems of which an expert with only a general training might miss".

We would add that, however good the scientific basis for a given composite indicator, its acceptance relies on negotiation.

2.1 Steps towards composite indicators

As first step towards the construction of a composite indicator, one should look at the indicators as an entity, with a view to investigate its structure. **Multivariate statistic** is a powerful tool to

[1] This Index is defined as a measure of the process of expanding people's capabilities (or choices) to function. In this case, composite indicators' use for advocacy is what makes them valuable.

achieve this objective. This type of analysis is, thereafter, of exploratory nature and is helpful in assessing the suitability of the dataset and providing an understanding of the implications of the methodological choices (e.g. weighting, aggregation) during the construction phase of the composite indicator. In the analysis, the statistical information inherent in the indicators' set can be dealt with grouping information along the two dimensions of the dataset, i.e. along indicators and along constituencies (e.g. countries, regions, sectors, etc.), not independently of each other.

Factor Analysis and Reliability/Item Analysis (e.g. Coefficient Cronbach Alpha) can be used to group the information on the indicators. The aim is to explore whether the different dimensions of the phenomenon are well balanced -from a statistical viewpoint- in the composite indicator. The higher the correlation between the indicators, the fewer statistical dimensions will be present in the dataset. However, if the statistical dimensions do not coincide with the theoretical dimensions of the dataset, then a revision of the set of the sub-indicators might be considered. Saisana et al. (2005) phrase that, depending on a school of thought, one may see a high correlation among indicators as something to correct for, e.g by making the weight for a given indicator inversely proportional to the arithmetic mean of the coefficients of determination for each bivariate correlation that includes the given indicator. On the other hand, practitioners of multi-criteria decision analysis would tend to consider the existence of correlations as a feature of the problem, not to be corrected for, as correlated indicators may indeed reflect non-compensable different aspects of the problem.

Cluster Analysis can be applied to group the information on constituencies (e.g. countries) in terms of their similarity with respect to the different sub-indicators. This type of analysis can serve multiple purposes, and it can be seen as:

(a) a purely statistical method of aggregation of the indicators,
(b) a diagnostic tool for assessing the impact of the methodological choices made during the construction phase of the composite indicator,
(c) a method of disseminating the information on the composite indicator, without losing the information on the dimensions of the indicators,
(d) a method for selecting groups of countries to impute missing data with a view to decrease the variance of the imputed values.

Clearly the ability of a composite to represent multidimensional concepts largely depends on the quality and accuracy of its components. **Missing data** are present in almost all composite indicators, and they can be missing either in a random or in a non-random fashion. However, there is often no basis upon which to judge whether data are missing at random or systematically, whilst most of the methods of imputation require a missing at random mechanism. When there are reasons to assume a non-random missing pattern, then this pattern must be explicitly modelled and included in the analysis. This could be very difficult and could imply ad hoc assumptions that are likely to deeply influence the result of the entire exercise.

Three generic approaches for dealing with missing data can be distinguished, i.e. case deletion, single imputation or multiple imputation. When an indicator is missing for a country, case deletion either removes the country from the analysis or the indicator from the analysis. The main disadvantage of case deletion is that it ignores possible systematic differences between complete and incomplete sample and may produce biased estimates if removed records are not a random sub-sample of the original sample. Furthermore, standard errors will, in general be larger in a reduced sample given that less information is used. The other two approaches see the missing data as part of the analysis and therefore try to impute values through either Single Imputation (e.g. Mean/Median/Mode substitution, Regression Imputation, Expectation-Maximisation

Imputation, etc.) or Multiple Imputation (e.g. Markov Chain Monte Carlo algorithm). The advantages of imputation include the minimisation of bias and the use of 'expensive to collect' data that would otherwise be discarded. In the words of Dempster and Rubin (1983): "The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can legitimately handled in this way and situations where standard estimators applied to real and imputed data have substantial bias."

Whenever indicators in a dataset are incommensurate with each other, and/or have different measurement units, it is necessary to bring these indicators to the same unit, to avoid adding up apples and pears. **Normalization** serves primarily this purpose. There are a number of normalization methods available, such as ranking, standardization, re-scaling, distance to reference country, categorical scales, cyclical indicators, balance of opinions. The selection of a suitable normalization method to apply to the problem at hand is not trivial and deserves special care. The normalization method should take into account the data properties and the objectives of the composite indicator. The issues that could guide the selection of the normalization method include: whether hard or soft data are available, whether exceptional behaviour needs to be rewarded/penalised, whether information on absolute levels matters, whether benchmarking against a reference country is requested, whether the variance in the indicators needs to be accounted for. For example, in the presence of extreme values, normalisation methods that are based on standard deviation or distance from the mean are preferred. Special care to the type of the normalisation method used needs to be given if the composite indicator values per country need to be comparable over time.

There is one further aspect which the normalization method may interfere with. This is the scale effect, i.e. the different measurement units in which an indicator can be expressed. Ebert and Welsch (2004) mention that particular attention needs to be placed if the raw data are expressed in different scales either interval scale (e.g. temperature in Celsius or Fahrenheit) or ratio scale (e.g. kilograms or pounds). In that case, a proper normalisation method should be applied to remove the scale effect from all indicators simultaneously. If for example, some indicators in the dataset are expressed on interval scale, whilst others on a ratio scale, then dividing by a reference value does not remove the scale effect from those indicators expressed on interval scale. However, the standardisation method does so.

Two types of transformations that are sometimes applied to the raw data prior to normalisation are truncation and functional form. The choice of trimming the tails of the indicators' distributions is supported by the need to avoid having extreme values overly dominate the result and, partially, to correct for data quality problems in such extreme cases. The functional transformation is applied to the raw data to represent the significance of marginal changes in its level. In most cases, the linear functional form is used on all of the variables, de facto. This approach is suitable if changes in the indicator's values are important in the same way, regardless of the level. If changes are more significant at lower levels of the indicator, the functional form should be concave down (e.g. log or the nth root). If changes are more important at higher levels of the indicator, the functional form should be concave up (e.g. exponential or power).

Central to the construction of a composite index is the need to combine in a meaningful way the different dimensions, which implies a decision on the **weighting** model and the aggregation procedure. Different weights may be assigned to indicators to reflect their economic significance (collection costs, coverage, reliability and economic reason), statistical adequacy, cyclical conformity, speed of available data, etc. Several weighting techniques are available, such as

weighting schemes based on statistical models (e.g. factor analysis, data envelopment analysis, unobserved components models), or on participatory methods (e.g. budget allocation, analytic hierarchy processes). For example, weights would be determined based on correlation coefficients or principal components analysis to overcome the "statistical" double counting problem when two or more indicators partially measure the same behaviour. Weights may also reflect the statistical quality of the data, thus higher weight could be assigned to statistically reliable data (data with low percentages of missing values, large coverage, sound values). In this case the concern is to reward only sub-indicators easy to measure and readily available, punishing the information that is more problematic to identify and measure. Indicators could also be weighted based on experts' opinion, who know policy priorities and theoretical backgrounds, to reflect the multiplicity of stakeholders' viewpoints. Weights usually have an important impact on the results of the composite indicator especially whenever higher weight is assigned to indicators on which some countries excel or fail. This is why weighting models need to be made explicit and transparent. Moreover, one should have in mind that, no matter which method is used, weights are essentially value judgments and have the property to make explicit the objectives underlying the construction of a composite (Rowena et al., 2004).

The issue of **aggregation** of the information conveyed by the different dimensions into a composite index comes together with the weighting. Different aggregation rules are possible. Sub-indicators could be summed up (e.g. linear aggregation), multiplied (geometric aggregation) or aggregated using non linear techniques (e.g. multi-criteria analysis). Each technique implies different assumptions and has specific consequences.

Linear aggregation can be applied when all indicators have the same measurement unit and further ambiguities related to the scale effects have been neutralized. Furthermore, linear aggregation implies full (and constant) compensability, i.e. poor performance in some indicators can be compensated by sufficiently high values of other indicators. Geometric aggregation is appropriate when strictly positive indicators are expressed in different ratio-scales, and it entails partial (non constant) compensability, i.e. compensability is lower when the composite indicator contains indicators with low values. The absence of synergy or conflict effects among the indicators is a necessary condition to admit either linear or geometric aggregation. Furthermore, linear aggregations reward sub-indicators proportionally to the weights, while geometric aggregations reward more those countries with higher scores. In both linear and geometric aggregations weights express trade-offs between indicators: the idea is that deficits in one dimension can be offset by surplus in another. However, when different goals are equally legitimate and important, then a non-compensatory logic may be necessary. This is usually the case when very different dimensions are involved in the composite, like in the case of environmental indexes, where physical, social and economic figures must be aggregated. If the analyst decides that an increase in economic performance can not compensate a loss in social cohesion or a worsening in environmental sustainability, then neither the linear nor the geometric aggregation are suitable. Instead, a non-compensatory multicriteria approach will assure non compensability by formalizing the idea of finding a compromise between two or more legitimate goals.

Doubts are often raised about the robustness of the results of the composite indicators and about the significance of the associated policy message. **Uncertainty analysis and sensitivity analysis** is a powerful combination of techniques to gain useful insights during the process of composite indicators building, including a contribution to the indicators' quality definition and an assessment of the reliability of countries' ranking.

As often noted, composite indicators may send misleading, non-robust policy messages if they are poorly constructed or misinterpreted. The construction of composite indicators involves stages where judgement has to be made. This introduces issues of uncertainty in the construction line of a composite indicator: selection of data, data quality, data editing (e.g. imputation), data normalisation, weighting scheme/weights, weights' values and aggregation method. All these sources of subjective judgement will affect the message brought by the composite indicator in a way that deserves analysis and corroboration. For example, changes in weights will almost in all cases lead to changes in rankings of countries. It is seldom that top performers becomes worse performance due to changes in weights but a change in ranking from for example ranking 2 to ranking 4 is not uncommon even in well-constructed composite indicators.

A combination of uncertainty and sensitivity analysis can help to gauge the robustness of the composite indicator, to increase its transparency and to help framing a debate around it. Uncertainty analysis (UA) focuses on how uncertainty in the input factors propagates through the structure of the composite indicator and affects the composite indicator values. Sensitivity analysis (SA) studies how much each individual source of uncertainty contributes to the output variance. In the field of building composite indicators, UA is more often adopted than SA (Jamison and Sandbu, 2001; Freudenberg, 2003) and the two types of analysis are almost always treated separately. A synergistic use of UA and SA is proven to be more powerful (Saisana et al., 2005; Tarantola *et al.*, 2000).

The types of questions for which an answer is sought via the application of UA&SA are:

(a) Does the use of one construction strategy versus another in building the composite indicator provide actually a partial picture of the countries' performance? In other words, how do the results of the composite indicator compare to a deterministic approach in building the composite indicator?
(b) How much do the uncertainties affect the results of a composite indicator with respect to a deterministic approach used in building the composite indicator?
(c) Which constituents (e.g. countries) have large uncertainty bounds in their rank (volatile countries) and therefore, if excluded, the stability of the system would increase?
(d) Which are the factors that affect the ranks of the volatile countries?

All things considered, a careful analysis of the uncertainties included in the development of a composite indicator can render the building of a composite indicator more robust. A plurality of methods (all with their implications) should be initially considered, because no model (construction path of the composite indicator) is a priori better than another, provided that internal coherence is always assured, as each model serves different interests. The composite indicator is no longer a magic number corresponding to crisp data treatment, weighting set or aggregation method, but reflects uncertainty and ambiguity in a more transparent and defensible fashion. The iterative use of uncertainty and sensitivity analysis during the development of a composite indicator can contribute to its well-structuring, provide information on whether the countries' ranking measures anything meaningful and could reduce the possibility that the composite indicator may send misleading or non-robust policy messages.

The **way of presenting composite indicators** is not a trivial issue. Composite indicators must be able to communicate the picture to decision-makers and users quickly and accurately. Visual models of these composite indicators must be able to provide signals, in particular, warning signals that flag for decision-makers those areas requiring policy intervention. The literature presents various ways for presenting the composite indicator results, ranging from simple forms,

such as tables, bar or line charts, to more sophisticated figures, such as the four-quadrant model (for sustainability), the Dashboard, etc.

If we were to stress the importance of visualising properly the composite indicators, we would use the general remark made by Shumpeter 1933:

"…as long as we are unable to put our arguments into figures, the voice of our science, although occasionally it may help to dispel gross errors, will never be heard by practical men."

One final suggestion for this introductory section concerns the '**Transparency**' of the indicator. It would be very useful, for developers, users and practitioners in general, if composite indicators could be made available via the web, along with the data, the weights and the documentation of the methodology. Given that composite indicators can be decomposed or disaggregated so as to introduce alternative data, weighting, normalisation approaches etc., the components of composites should be available electronically as to allow users to change variables, weights, etc. and to replicate sensitivity tests.

2.1 Requirements for quality control

As mentioned above the concept of **quality of the composite indicators** is not only a function of the quality of its underlying data (in terms of relevance, accuracy, credibility, etc.) but also of the quality of the methodological process used to build the composite indicator itself[2]. The safe use of the composite requires proper evidence that the composite will provide reliable results. If the user simply does not know, or is not sure about the testing and certification of the composite, then composite's quality is low. Up to now, tests for the quality of quantitative information have been much undeveloped. There are statistical hypothesis tests, and elaborated formal theories of decision-making, but none of these approaches helps with the simple question that a decision-maker wants to ask: is this message reliable, can I use it safely?

A notational system called NUSAP (an acronym for five categories: Numeral, Unit, Spread, Assessment, Pedigree) has been devised to characterise the quality of quantitative information based in large part on the experience of research work in the matured natural sciences (Funtowicz and Ravetz, 1990).

The categorical scheme on which NUSAP is based enables providers and users of composite indicators to communicate their quality. One category of NUSAP, the pedigree, is an evaluative description of the procedure used to build the composite indicator. The pedigree is expressed by means of a matrix Each column of the matrix represents one phase of the construction process. For example, the first phase of the process could be "problem definition and purpose". A score is assigned to each phase according to the mode the phase itself has been executed. In the example, the phase "problem definition and purpose" could be executed in various modes: "result of negotiation", "purely science-based", "based on different subjective interpretations", "purely abstract" or "not explored". In very general terms, the pedigree is laid out as in Table 2.1, where the top row has grade 4 and the bottom two rows, 0. For a numerical evaluation, average scores of 4 downwards are rated as High, Good, Medium, Low and Poor. All the scores are then elaborated to provide an assessment of the quality of the process, which in turns suggests recursive actions for the improvement of the process itself.

[2] This chapter is based on text available on www.nusap.net

The whole pedigree matrix is conditioned by the principle that statistical work is a highly articulated social activity. Thus, the pedigree matrix, with its multiplicity of categories, enables a considerable variety of evaluative descriptions of the composite indicator to be simply scored and coded. In practical cases, a specific pedigree matrix has to be constructed for each specific composite indicator. An example of pedigree matrix used to characterise the quality of a set of statistical indicators of knowledge economy can be found in Sajeva, 2004. The pedigree matrix builds on a series of interviews made to statisticians, concerning the process they followed for the development of the indicators (the complete text of one such interview is reported in Sajeva, 2004).

Table 2.1 *The Pedigree Matrix for Statistical Information*

Grade	Definitions & Standards	Data-collection & Analysis	Institutional Culture	Review
4	Negotiation	Task-force	Dialogue	External
3	Science	Direct Survey	Accommodation	Independent
2	Convenience	Indirect Survey	Obedience	Regular
1	Symbolism	Educated Guess	Evasion	Occasional
0	Inertia	Fiat	No-contact	None
0	Unknown	Unknown	Unknown	Unknown

In the following Sections we present a detailed discussion of some of the main steps in the construction of a composite indicator.

3. Multivariate analysis

The information inherent in a dataset of sub-indicators that measure the performance of several countries can be studied along two dimensions, i.e. along sub-indicators and along countries, not independently of each other.

Information on sub-indicators. The analyst must first decide whether the nested structure of the composite indicator is well defined and if the set of available sub-indicators is sufficient or appropriate to describe the unknown phenomenon. This decision can be based both on experts' opinion (e.g. experts in the relevant field will tell which indicators better capture the sustainability or the quality of the education) and on the statistical structure of the dataset. Factor Analysis and Reliability/Item Analysis can be used complementarily to explore whether the different dimensions of the phenomenon are well balanced -from a statistical viewpoint- in the composite indicator. If this is not true, a revision of the set of the sub-indicators might be considered. For instance, in the e-business readiness index the human capital factor is clearly understated, whilst the technological factor is favoured. In the same example, the distinction between "use" and "adoption" of information and communication technologies is not supported statistically, since Principal Components Analysis shows that some of the sub-indicators conceptually allocated to "use" are better associated with the sub-indicators on "adoption".

Information on countries. The use of cluster analysis to group countries in terms of similarity between different sub-indicators can serve as:

(e) a purely statistical method of aggregation,

(f) a diagnostic tool for assessing the impact of the methodological choices made during the construction phase of the composite indicator,
(g) a method of disseminating the information on the composite indicator, without losing the information on the dimensions of the sub-indicators,
(h) a method for selecting groups of countries to impute missing data with a view to decrease the variance of the imputed values.

Cluster Analysis could, thereafter, be useful in different sections of this document.

The notation that we will adopt throughout this document is the following.

$x_{q,c}^{t}$: raw value of sub-indicator q for country c at time t, with $q=1,...,Q$ and $c=1,...,M$

$I_{q,c}^{t}$: normalised value of sub-indicator

$w_{r,q}$: weight associated to sub-indicator q, with $r=1,...,R$

CI_{c}^{t} : value of the composite indicator for country c at time t.

Note that time suffix is present only in Section 5. For reasons of clarity the time suffix has been dropped out. When no time indication is present, the reader should consider that all variables have the same time dimension. The rest of the notation will be introduced section by section.

3.1 Grouping Information on sub-indicators

3.1.1 Principal Components Analysis

The goal of the Principal Components Analysis (PCA) is to reveal how different variables change in relation to each other, or how they are associated. This is achieved by transforming correlated original variables into a new set of uncorrelated variables using the covariance matrix, or its standardized form – the correlation matrix. The new variables are linear combinations of the original ones and are sorted into descending order according to the amount of variance they account for in the original set of variables. Usually correlations among original variables are large enough so that the first few new variables, termed ***principal components*** account for most of the variance in the dataset. If this holds, no essential insight is lost by further analysis or decision making, and parsimony and clarity in the structure of the relationships are achieved. A brief description of the PCA approach is provided in the next paragraphs. For a detailed discussion on the PCA the reader is referred to Jolliffe (1986), Jackson (1991) and Manly (1994). Social scientists may also find the shorter monograph by Dunteman (1989) to be helpful.

The technique of PCA was first described by Karl Pearson in 1901. A description of practical computing methods came much later from Hotelling in 1933. The objective of the analysis is to take Q variables $x_1, x_2, ...x_Q$ and find linear combinations of these to produce principal components $Z_1, Z_2, ...Z_Q$ that are uncorrelated, following

$$\begin{aligned} Z_1 &= a_{11}x_1 + a_{12}x_2 + ... + a_{1Q}x_Q \\ Z_2 &= a_{21}x_1 + a_{22}x_2 + ... + a_{2Q}x_Q \\ &... \\ Z_Q &= a_{Q1}x_1 + a_{Q2}x_2 + ... + a_{QQ}x_Q \end{aligned} \qquad (3.1)$$

At this point there are still Q principal components, i.e. as many as there are variables. The next step is to select the first, say $P<Q$ principal components that preserve a "*high"* amount of the cumulative variance of the original data.
The lack of correlation in the principal components is a useful property because it means that the principal components are measuring different "statistical dimensions" in the data. When the objective of the analysis is to present a huge data set using a few variables then in applying PCA there is the hope that some degree of economy can be achieved if the variation in the Q original x variables can be accounted for by a small number of Z variables. It must be stressed that PCA cannot always reduce a large number of original variables to a small number of transformed variables. Indeed, if the original variables are uncorrelated then the analysis does absolutely nothing. On the other hand, a significant reduction is obtained when the original variables are highly correlated, positively or negatively.

The weights a_{ij} applied to the variables x_j in Equation (3.1) are chosen so that the principal components Z_i satisfy the following conditions:

(i) they are uncorrelated (orthogonal),
(ii) the first principal component accounts for the maximum possible proportion of the variance of the set of x s, the second principal component accounts for the maximum of the remaining

variance and so on until the last of the principal component absorbs all the remaining variance not accounted for by the preceding components, and[3]

(iii) $\alpha_{i1}^2 + \alpha_{i2}^2 + ... + \alpha_{iQ}^2 = 1, i = 1,2,...,Q$

In brief, PCA just involves finding the ***eigenvalues*** λ_j of the sample covariance matrix *CM*,

$$CM = \begin{bmatrix} cm_{11}\ cm_{12}\ ...cm_{1Q} \\ cm_{21}\ cm_{22}\ ...cm_{2Q} \\ ... \\ cm_{Q1}\ cm_{Q2}\ ...cm_{QQ} \end{bmatrix} \quad (3.2)$$

where the diagonal element cm_{ii} is the variance of x_i and cm_{ij} is the covariance of variables x_i and x_j. The eigenvalues of the matrix *CM* are the variances of the principal components. There are Q eigenvalues, some of which may be negligible. Negative eigenvalues are not possible for a covariance matrix. An important property of the eigenvalues is that they add up to the sum of the diagonal elements of CM. This means that the sum of the variances of the principal components is equal to the sum of the variances of the original variables,

$$\lambda_1 + \lambda_2 + ... + \lambda_Q = cm_{11} + cm_{22} + ... + cm_{QQ} \quad (3.3)$$

In order to avoid one variable having an undue influence on the principal components it is common to standardize the variables x*s* to have zero means and unit variances at the start of the analysis. The matrix CM then takes the form of the correlation matrix (Table 3.1). For the TAI example, the highest correlation is found between the sub-indicators ELECTRICITY & INTERNET with a coefficient of 0.84.

Table 3.1. *Correlation matrix for the TAI sub-indicators, n=23. Marked correlations are statistically significant at p < 0.05.*

	PATENTS	ROYALTIES	INTERNET	EXPORTS	TELEPHONES	ELECTRICITY	SCHOOLING	ENROLMENT
PATENTS	1.00	0.13	-0.09	**0.45**	0.28	0.03	0.22	0.08
ROYALTIES		1.00	**0.46**	0.25	**0.56**	0.32	0.30	0.06
INTERNET			1.00	**-0.45**	**0.56**	**0.84**	**0.63**	0.27
EXPORTS				1.00	0.00	-0.36	-0.35	-0.03
TELEPHONES					1.00	**0.64**	0.30	0.33
ELECTRICITY						1.00	**0.65**	0.26
SCHOOLING							1.00	0.08
ENROLMENT								1.00

[3] For reasons of clarity in this section we substitute the indexing $q=1,...Q$ with the indexing $i=1,...,Q$ and $j=1,...,Q$.

Table 3.2 gives the eigenvalues of the correlation matrix of the eight sub-indicators (standardised values) that compose TAI. Note that the sum of the eigenvalues is equal to the number of sub-indicators ($Q = 8$). Figure 3.1 (left) is a graphical presentation of the eigenvalues in descending order. Given that the correlation matrix rather than the covariance matrix is used in the PCA, all 8 sub-indicators are assigned equal weights in forming the principal components (Chatfield and Collins, 1980). The first Principal Component explains the maximum variance in all the sub-indicators – eigenvalue of 3.3. The second principal component explains the maximum amount of the remaining variance – a variance of 1.7. The third and fourth principal components have an eigenvalue close to 1. The last four principal components explain the remaining 12.8% of the variance in the dataset.

Table 3.2. *Eigenvalues of the 8 sub-indicators' set in TAI (n=23). Extraction method: Principal Components Analysis*

	Eigenvalue	% of variance	Cumulative %
1	3.3	41.9	41.9
2	1.7	21.8	63.7
3	1.0	12.3	76.0
4	0.9	11.1	87.2
5	0.5	6.0	93.2
6	0.3	3.7	96.9
7	0.2	2.2	99.1
8	0.1	0.9	100.0

A drawback of the conventional PCA is that it does not allow for inference on the properties of the general population. This is because, traditionally, drawing such inferences requires certain distributional assumptions to be made regarding the population characteristics, and the PCA techniques are not based upon such assumptions (see below on the "Assumptions of the PCA"). Furthermore, in a traditional PCA framework, there is no estimation of the statistical precision of the results, which is essential for relatively small sample sizes as in the present case of the TAI example. Therefore, the bootstrap method has been proposed to be utilized in conjuction with PCA to make inferences about the population (Efron and Tibshirani, 1991, 1993). Bootstrap refers to the process of randomly re-sampling the original data set to generate new data sets. Estimates of the relevant statistics are made for each bootstrap sample. A very large number of bootstrap samples will give satisfactory results but the computation may be cumbersome. Various values have been suggested, ranging from 25 (Efron and Tibshirani, 1991) to as high as 1000 (Efron, 1987; Mehlman et al., 1995).

An issue that arises at this point is whether the TAI data set for the 23 countries can be viewed as a 'random' sample of the entire population as required by the bootstrap procedures (Efron 1987; Efron and Tibshirani 1993). Several points can be made regarding the issues of randomness and representativeness of the data. First, it is often difficult to obtain complete information for a data set in the social sciences because, unlike the natural sciences, controlled experiments are not always possible, as in the case here. As Efron and Tibshirani (1993) state: 'in practice the selection process is seldom this neat […], but the conceptual framework of random sampling is still useful for understanding statistical inferences.' Second, the countries included in the restricted set show no apparent pattern as to whether or not they are predominately developed or developing countries. In addition, the countries of varying sizes span all the major continents of the world, ensuring a wide representation of the global state of technological development. Consequently, the restricted set could be considered as representative of the total population. A

third point on the data quality is that a certain amount of measurement error is likely to exist. While such measurement error can only be controlled at the data collection stage, rather than at the analytical stage, it is argued that the data represent the best estimates currently available (United Nations, 2001, p. 46).

Figure 3.1 (right) demonstrates graphically the relationship between the eigenvalues from the deterministic PCA, their bootstrapped confidence intervals (5^{th} and 95^{th} percentiles) and the ranked principal components. These confidence intervals allow one to generalize the conclusions concerning the small set of the sub-indicators (23 countries) to the entire population (e.g. of 72 countries or even more general), rather than confining the conclusions only to the sample set being analyzed. Bootstrapping has been performed for 1000 sample sets of size 23 (random sampling with replacement). It is shown that the values of the eigenvalues drop sharply at the beginning and then gradually approach zero after a certain point.

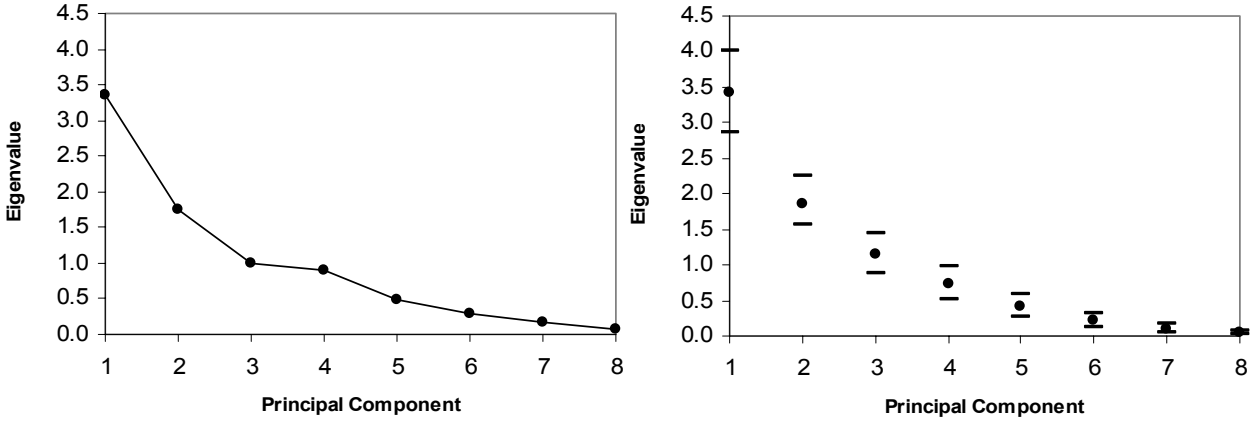


Figure 3.1. *Eigenvalues for the 8 sub-indicators in the TAI examples (23 countries). Eigenvalues from traditional Principal Components Analysis - Scree plot (left graph), Bootstrapped eigenvalues, 1000 samples randomly selected with replacement (right graph).*

The correlation coefficients between the principal components Z and the variables x are called ***component loadings***, $r(Z_j, x_i)$. In case of uncorrelated variables *x*, the loadings are equal to the weights a_{ij} given in equation (3.1). Analogous to Pearson's r, the squared loading is the percent of variance in that variable explained by the principal component. The ***component scores*** are the scores of each case (country in our example) on each principal component. The component score for a given case for a principal component is calculated by taking the case's standardized value on each variable, multiplying by the corresponding loading of the variable for the given principal component factor, and summing these products.

Table 3.3 presents the components loadings for the TAI sub-indicators. High and moderate loadings (>0.50) indicate how the sub-indicators are related to the principal components. It can be seen that with the exception of PATENTS and ROYALTIES, all the other sub-indicators are entirely accounted for by one principal component alone and that the high and moderate loadings are all found in the first four principal components. An undesirable property of these components is that two sub-indicators are related strongly to two principal components.

Table 3.3. *Component loadings for the TAI example (23 countries) of the eight sub-indicators. Extraction method: principal components. Loadings greater than 0.5 (absolute values) are highlighted.*

	1	2	3	4	5	6	7	8
PATENTS	-0.11	**-0.75**	0.13	**0.60**	-0.10	-0.12	-0.17	0.05
ROYALTIES	**-0.56**	-0.48	0.22	**-0.54**	0.27	-0.17	-0.04	0.10
INTERNET	**-0.92**	0.21	0.02	-0.10	0.04	0.11	-0.27	-0.13
EXPORTS	0.35	**-0.85**	0.01	-0.13	0.11	0.35	0.06	-0.08
TELEPHONES	**-0.76**	-0.39	-0.16	-0.16	-0.41	-0.16	0.16	-0.09
ELECTRICITY	**-0.91**	0.13	0.01	0.07	-0.19	0.30	0.04	0.16
SCHOOLING	**-0.74**	0.11	0.37	0.39	0.33	-0.02	0.20	-0.07
ENROLMENT	-0.36	-0.12	**-0.87**	0.15	0.26	-0.03	0.02	0.02

The question of how many principal components should be retained in the analysis without losing too much information and how the interpretation of the components might be improved are addressed without further ado in the following section on Factor Analysis.

3.1.2 Factor Analysis

Factor analysis (FA) has similar aims to PCA. The basic idea is still that it may be possible to describe a set of Q variables $x_1, x_2, ..., x_Q$ in terms of a smaller number of m factors, and hence elucidate the relationship between these variables. There is however, one important difference: PCA is not based on any particular statistical model, but FA is based on a rather special model (Spearman, 1904).
In a general form this model is given by:

$$\begin{aligned} x_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + ... + \alpha_{1m}F_m + e_1 \\ x_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + ... + \alpha_{2m}F_m + e_2 \\ &... \\ x_Q &= \alpha_{Q1}F_1 + \alpha_{Q2}F_2 + ... + \alpha_{Qm}F_m + e_Q \end{aligned} \qquad (3.4)$$

where x_i is a variable with zero mean and unit variance; $\alpha_{i1}, \alpha_{i2}, ..., \alpha_{im}$ are the factor loadings related to the variable X_i; $F_1, F_2,...,F_m$ are m uncorrelated common factors, each with zero mean and unit variance; and e_i are the Q specific factors supposed independently and identically distributed with zero mean. There are several approaches to deal with the model in equation (3.4), e.g. communalities, maximum likelihood factors, centroid method, principal axis method, etc. All them giving different values for the factos. The most common is the use of PCA to extract the first m principal components and consider them as factors and neglect the remaining. Principal components factor analysis is most preferred in the development of composite indicators (see Section 6), e.g. Product Market Regulation Index (Nicoletti et al. 2000), as it has the virtue of simplicity and allows the construction of weights representing the information content of sub-indicators. Notice however that different extraction methods supply different values for the factors thus for the weights, influencing the score of the composite and the corresponding country ranking.

On the issue of how factors should be retained in the analysis without losing too much information methodologists' opinions differ. The decision of when to stop extracting factors basically depends on when there is only very little "random" variability left, and it is rather arbitrary. However, various guidelines ("stopping rules") have been developed, and they are

reviewed below, roughly in the order of frequency of their use in social science (see Dunteman, 1989: 22-3).

- **Kaiser criterion.** Drop all factors with eigenvalues below 1.0. The simplest justification to this rule is that it doesn't make sense to add a factor that explains less variance than is contained in one sub-indicator. According to this rule, 3 factors should be retained in the analysis of the TAI example, although the 4th factor follows closely with an eigenvalues of 0.90 (see Table 3.2).
- **Scree plot.** This method proposed by Cattell plots the successive eigenvalues, which drop off sharply and then tend to level off. It suggests retaining all eigenvalues in the sharp descent before the first one on the line where they start to level off. This approach would result in retaining 3 factors in the TAI example (Figure 3.1).
- **Variance explained criteria.** Some researchers simply use the rule of keeping enough factors to account for 90% (sometimes 80%) of the variation. The first 4 factors account for 87.2% of the total variance (see Table 3.2).
- **Joliffe criterion.** Drop all factors with eigenvalues under 0.70. This rule may result in twice as many factors as the Kaiser criterion, and it is less often used. In the present case study, this criterion would have lead to the selection of 4 factors.
- **Comprehensibility**. Though not a strictly mathematical criterion, there is much to be said for limiting the number of factors to those whose dimension of meaning is readily comprehensible. Often this is the first two or three.
- A relatively recent method for deciding on the number of factors to retain combines the **bootstrapped eigenvalues and eigenvectors** (Jackson 1993, Yu et al. 1998). Based on a combination of the Kaiser criterion and the bootstrapped eigenvalues, we should consider the first 4 factors in the TAI example.

In light of the above analysis, we retain the first four principal components as identified by the bootstrap eigenvalue approach combined with the Kaiser criterion. This choice implies a greater willingness to overstate the significance of the fourth component and be in line with the idea that there are four main categories of technology achievement indicators.

After choosing the number of factors to keep, ***rotation*** is a standard step performed to enhance the interpretability of the results (see for instance Kline, 1994). With rotation the sum of eigenvalues is not affected by rotation, but rotation, changing the axes, will alter the eigenvalues of particular factors and will change the factor loadings. There are various rotational strategies that have been proposed. The goal of all of these strategies is to obtain a clear pattern of loadings. However, different rotations imply different loadings, and thus different meanings of principal components - a problem some cite as a drawback to the method. The most common rotation method is the "varimax rotation".

Table 3.4 presents the factor loadings for the first factors in the TAI example. Note that the eigenvalues have been affected by the rotation. The variance accounted for by the rotated components is spread more evenly than for the unrotated components (Table 3.2). The first four factors account now for 87% of the total variance and are not sorted into descending order according to the amount of the original's dataset variance explained. The first factor has high positive coefficients (loadings) with INTERNET (0.79), ELECTRICITY (0.82) and SCHOOLING (0.88). Factor 2 is mainly dominated by PATENTS and EXPORTS, whilst ENROLMENT is exclusively loaded on Factor 3. Finally, Factor 4 is formed by ROYALTIES and TELEPHONES. Yet, despite the rotation of factors, the sub-indicator of EXPORTS has

sizeable loadings in both Factor 1 (negative loading) and Factor 2 (positive loading). A meaningful interpretation of the factors is not straightforward. Furthermore, the statistical treatment of the eight sub-indicators results in different groups (factors) than the conceptual ones (see Table A.1 in Appendix).

Table 3.4. *Rotated factor loadings for the TAI example (23 countries) of the eight sub-indicators. Extraction method: principal components, varimax normalised rotation. Positive loadings greater than 0.5 are highlighted.*

	Factor 1	Factor 2	Factor 3	Factor 4
PATENTS	0.07	**0.97**	0.06	0.06
ROYALTIES	0.13	0.07	-0.07	**0.93**
INTERNET	**0.79**	-0.21	0.21	0.42
EXPORTS	-0.64	**0.56**	-0.04	0.36
TELEPHONES	0.37	0.17	0.38	**0.68**
ELECTRICITY	**0.82**	-0.04	0.25	0.35
SCHOOLING	**0.88**	0.23	-0.09	0.09
ENROLMENT	0.08	0.04	**0.96**	0.04
Explained variance	2.64	1.39	1.19	1.76
Cumulative variance explained (%)	33	50	65	87

Another method of extracting factors that deals with the uncorrelation issue of the specific factors would have given different results. Just to give an example, Table 3.5 presents the rotated factor loadings of the four factors for the TAI case study (extraction method: principal factors maximum likelihood). For instance, ELECTRICITY and SCHOOLING are not loaded any more both on F1, but ELECTRICITY is loaded on F4 and SCHOOLING on F3. There is 76% variance that is common in the sub-indicators set and expressed by the four rotated common factors. In contrast, the total variance explained in the previous analysis by the four rotated principal components was much higher (87%).

Table 3.5. *Rotated factor loadings for the TAI example (23 countries). Extraction method: principal factors maximum likelihood, varimax normalised rotation.*

	Factor 1	Factor 2	Factor 3	Factor 4
PATENTS	0.01	0.11	**0.88**	0.13
ROYALTIES	**0.96**	0.14	0.09	0.18
INTERNET	0.31	**0.56**	-0.29	**0.60**
EXPORTS	0.29	-0.45	**0.58**	-0.14
TELEPHONES	0.41	0.13	0.18	**0.73**
ELECTRICITY	0.13	0.57	-0.13	**0.73**
SCHOOLING	0.14	**0.95**	0.10	0.14
ENROLMENT	-0.01	0.03	0.03	0.39
Explained Variance	1.31	1.80	1.27	1.67
Cumulative variance explained (%)	16	39	55	76

To sum up the steps of PCA/FA as exploratory analysis method:
1. Calculate the covariance/correlation matrix: if the correlations between sub-indicators are small, it is unlikely that they share common factors.
2. Identify the number of factors that are necessary to represent the data and the method for calculating them.

3. Rotate factors to enhance their interpretability (by maximizing loading of sub-indicators individual factors).

There are several assumptions in the application of PCA/FA, which we are discussed in the box below. These assumptions are mentioned in almost all textbooks, yet they are often neglected when composite indicators are developed.

Box: Assumptions in Principal Components Analysis and Factor Analysis

1. **Enough number of cases**. The question of how many cases (or countries) are necessary to do PCA/FA has no scientific answer and methodologists' opinions differ. Alternative arbitrary rules of thumb in descending order of popularity include those below.

   (a) Rule of 10. There should be at least 10 cases for each variable.
   (b) 3:1 ratio. The cases-to-variables ratio should be no lower than 3 (Grossman et al. 1991).
   (c) 5:1 ratio. The cases-to-variables ratio should be no lower than 5 (Bryant and Yarnold, 1995; Nunnaly 1978, Gorsuch 1983).
   (d) Rule of 100: The number of cases should be the larger between (5 × number of variables), and 100. (Hatcher, 1994).
   (e) Rule of 150: Hutcheson and Sofroniou (1999) recommend at least 150 - 300 cases, more toward 150 when there are a few highly correlated variables.
   (f) Rule of 200. There should be at least 200 cases, regardless of the cases-to-variables ratio (Gorsuch, 1983).
   (g) Significance rule. There should be 51 more cases than the number of variables, to support chi-square testing (Lawley and Maxwell, 1971)

   These rules are not mutually exclusive. Bryant and Yarnold (1995), for instance, endorse both the cases-to-variables ratio and the Rule of 200. In the TAI example, there are 23:8 cases-to-variables, therefore the first and the second rule are satisfied.

2. **No bias in selecting sub-indicators**. The exclusion of relevant sub-indicators and the inclusion of irrelevant sub-indicators in the correlation matrix being factored will affect, often substantially, the factors which are uncovered. Although social scientists may be attracted to factor analysis as a way of exploring data whose structure is unknown, knowing the factorial structure in advance helps select the sub-indicators to be included and yields the best analysis of factors. This dilemma creates a chicken-and-egg problem. Note this is not just a matter of including all relevant sub-indicators. Also, if one deletes sub-indicators arbitrarily in order to have a "cleaner" factorial solution, erroneous conclusions about the factor structure will result (see Kim and Mueller, 1978a: 67-8).

3. **No outliers**. As with most techniques, the presence of outliers can affect interpretations arising from PCA/FA. One may use Mahalanobis distance to identify cases which are multivariate outliers and remove them prior to the analysis. Alternatively, one can also create a dummy variable set to 1 for cases with high Mahalanobis distance, then regress this dummy on all other variables. If this regression is non-significant (or simply has a low R-squared for large samples) then the outliers are judged to be at random and there is less danger in retaining them. The ratio of the regression coefficients indicates which variables are most associated with the outlier cases.

4. **Assumption of interval data**. Kim and Mueller (1978b, pp.74-75) note that ordinal data may be used if it is thought that the assignment of ordinal categories to the data does not seriously

distort the underlying metric scaling. Likewise, these authors allow the use of dichotomous data if the underlying metric correlations between the variables are thought to be moderate (.7) or lower. The result of using ordinal data is that the factors may be much harder to interpret. Note that categorical variables with similar splits will necessarily tend to correlate with each other, regardless of their content (see Gorsuch, 1983). This is particularly apt to occur when dichotomies are used. The correlation will reflect similarity of "difficulty" for items in a testing context, hence such correlated variables are called *difficulty factors*. The researcher should examine the factor loadings of categorical variables with care to assess whether common loading reflects a difficulty factor or substantive correlation.

5. **Linearity**. Principal components factor analysis (PFA), which is the most common variant of FA, is a linear procedure. Of course, as with multiple linear regression, nonlinear transformation of selected variables may be a pre-processing step, but this is not common. The smaller the sample size, the more important it is to screen data for linearity.

6. **Multivariate normality** of data is required for related significance tests. PCA and PFA have no distributional assumptions. Note, however, that a variant of factor analysis, maximum likelihood factor analysis, does assume multivariate normality. The smaller the sample size, the more important it is to screen data for normality. Moreover, as factor analysis is based on correlation (or sometimes covariance), both correlation and covariance will be attenuated when variables come from different underlying distributions (ex., a normal vs. a bimodal variable will correlate less than 1.0 even when both series are perfectly co-ordered).

7. **Underlying dimensions** shared by clusters of sub-indicators are assumed. If this assumption is not met, the "garbage in, garbage out" principle applies. Factor analysis cannot create valid dimensions (factors) if none exist in the input data. In such cases, factors generated by the factor analysis algorithm will not be comprehensible. Likewise, the inclusion of multiple definitionally-similar sub-indicators representing essentially the same data will lead to tautological results.

8. **Strong intercorrelations** are not mathematically required, but applying factor analysis to a correlation matrix with only low intercorrelations will require for solution nearly as many factors as there are original variables, thereby defeating the data reduction purposes of factor analysis. On the other hand, too high inter-correlations may indicate a multi-collinearity problem and collinear terms should be combined or otherwise eliminated prior to factor analysis.

   (a) The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is a statistics for comparing the magnitudes of the observed correlation coefficients to the magnitudes of the partial correlation coefficients. The concept is that the partial correlations should not be very large if one is to expect distinct factors to emerge from factor analysis (see Hutcheson and Sofroniou, 1999, p.224). A KMO statistic is computed for each individual sub-indicator, and their sum is the KMO overall statistic. KMO varies from 0 to 1.0. A KMO overall should be .60 or higher to proceed with factor analysis (Kaiser and Rice, 1974), though realistically it should exceed 0.80 if the results of the principal components analysis are to be reliable. If not, it is recommended to drop the sub-indicators with the lowest individual KMO statistic values, until KMO overall rises above .60.

   (b) Variance-inflation factor (VIF) is simply the reciprocal of tolerance. A VIF value greater than 4.0 is an arbitrary but common cut-off criterion for suggesting that there is a multi-

collinearity problem. Some researchers use the more lenient cutoff VIF value of 5.0.

(c) The Bartlett's test of sphericity is used to test the null hypothesis that the sub-indicators in a correlation matrix are uncorrelated, that is to say that the correlation matrix is an identity matrix. The statistic is based on a chi-squared transformation of the determinant of the correlation matrix. However, as Bartlett's test is highly sensitive to sample size (Knapp and Swoyer 1967), Tabachnick and Fidell (1989, p.604) suggest implementing it with the KMO measure (point a above).

PCA/FA as exploratory analysis

Advantages	**Disadvantages**
▪ Can summarise a set of sub-indicators while preserving the maximum possible proportion of the total variation in the original set. ▪ Largest factor loadings are assigned to the sub-indicators that have the largest variation across countries (a desirable property for cross-country comparisons, as sub-indicators that are similar across countries are of little interest and cannot possibly explain differences in performance)	▪ Correlations do not necessarily represent the *real influence* of the sub-indicators on the phenomenon being measured. ▪ Sensitive to modifications in the basic data: data revisions and updates (e.g. new countries). ▪ Sensitive to the presence of outliers, which may introduce a spurious variability in the data. ▪ Sensitive to small-sample problems, which are particularly relevant when the focus is on a limited set of countries. ▪ Minimisation of the contribution of sub-indicators which do not move with other sub-indicators.

Examples of use
Environmental Sustainability Index
General Indicator of Science & Technology
Internal Market Index
Business Climate Indicator

3.1.3 Cronbach Coefficient Alpha

A way to investigate the degree of the correlations among a set of variables is to use the Cronbach Coefficient Alpha, c-alpha hencefort, (Cronbach, 1951). The c-alpha is the most common estimate of internal consistency of items in a model or survey – Reliability/Item Analysis (e.g. Boscarino et al., 2004; Raykov, 1998; Cortina, 1993; Feldt et al., 1987; Green et al., 1977; Hattie, 1985; Miller, 1995). It assesses how well a set of items (in our terminology sub-indicators) measures a single unidimensional object (e.g. attitude, phenomenon etc.).

Cronbach's Coefficient Alpha can be defined as:

$$\alpha_c = \left(\frac{Q}{Q-1}\right)\frac{\sum_{i\neq j} cov(x_i, x_j)}{var(x_o)} = \left(\frac{Q}{Q-1}\right)\left(1 - \frac{\sum_j var(x_j)}{var(x_o)}\right) \quad c = 1,..,M;\ i,j = 1,..,Q \qquad (3.5)$$

where M as usual indicates the number of countries considered, Q the number of sub-indicators available, and $x_o = \sum_{q=1}^{Q} x_j$ is the sum of all sub-indicators. C-alpha measures the portion of total variability of the sample of sub-indicators due to the correlation of indicators. It grows with the number of sub-indicators and with the covariance of each pair of them. If no correlation exists and sub-indicators are independent then c-alpha is equal to zero, while if sub-indicators are perfectly correlated the c-alpha is equal to one.

C-alpha is not a statistical test but a coefficient of reliability based on the correlations between sub-indicators: if the correlation of high, then there is evidence that the sub-indicators are measuring the same underlying construct. Therefore a high c-alpha, or equivalently a high "reliability", means that the sub-indicators considered measure well the latent phenomenon. Though widely interpreted as such, strictly speaking c-alpha is *not a measure of unidimensionality*. A set of sub-indicators can have a high alpha and still be multidimensional. This happens when there are separate clusters of sub-indicators (separate dimensions) which inter-correlate highly, even though the clusters themselves are not highly correlated. An issue is how large the c-alpha must be. Nunnally (1978) suggests 0.7 as an acceptable reliability threshold. Yet, some authors use 0.75 or 0.80 as cut-off value, while others are as lenient as 0.60. In general this varies by discipline.

If the variances of the sub-indicators vary widely, like in our test case, a standard practice is to standardize the sub-indicators to a standard deviation of 1 before computing the coefficient alpha. In our notation this would mean substituting x_i with I_i. The c-alpha is 0.70 for the dataset of the 23 countries, which is equal to the Nunnally's cutoff value. An interesting exercise is to determine how the c-alpha varies with the deletion of each sub-indicator at a time. This helps to detect the existence of clusters of sub-indicators, thus it is useful to determine the nested structure of the composite. If the reliability coefficient increases after deleting a sub-indicator from the scale, one can assume that the sub-indicator is not correlated highly with other sub-indicators in the scale.

Table 3.6 presents the values for the Cronbach coefficient alpha and the correlation with the total after deleting one sub-indicator at-a-time. TELEPHONES has the highest variable-total correlation and if deleted the coefficient alpha would be as low as 0.60. If EXPORTS were to be deleted from the set then the value of standardized coefficient alpha will increase from the current 0.70 to 0.77. Note that the same sub-indicator has the lowest variable-total correlation value (-0.108). This indicates that EXPORTS is not measuring the same construct as the rest of the sub-indicators are measuring. Note also, that the factor analysis in the previous section had indicated ENROLMENT as the sub-indicator that shares the least amount of common variance with the other sub-indicators. Although both factor analysis and the Cronbach coefficient alpha are based on correlations among sub-indicators, their conceptual framework is different.

Table 3.6. *Cronbach coefficient alpha results for the 23 countries after deleting one sub-indicator (standardised values) at-a-time*

Deleted sub-indicator	Correlation with total	Cronbach coefficient alpha
PATENTS	0.261	0.704
ROYALTIES	0.527	0.645
INTERNET	0.566	0.636
EXPORTS	-0.108	0.774
TELEPHONES	0.701	0.603
ELECTRICITY	0.614	0.624
SCHOOLING	0.451	0.662
ENROLMENT	0.249	0.706

Coefficient Cronbach alpha

Advantages

- It measures the internal consistency in the set of sub-indicators, i.e. how well they describe a unidimensional construct. Thus it is useful to cluster similar objects.

Disadvantages

- Correlations do not necessarily represent the real influence of the sub-indicators on the phenomenon expressed by the composite indicator.
- Cronbach coefficient alpha is meaningful only when the composite indicator is computed as a 'scale' (i.e. as the sum of the sub-indicators).

Examples of use

Compassion Fatigue (Boscarino et al., 2004)
Secondary trauma (Boscarino et al., 2004)
Job burnout (Boscarino et al., 2004)
Success of software process implementation

3.2 Grouping information on countries

3.2.1 Cluster analysis

Cluster analysis (CLA) is the name given to a collection of algorithms used to classify objects, e.g. countries, species, individuals (Anderberg 1973, Massart and Kaufman 1983). The classification has the aim of reducing the dimensionality of a dataset by exploiting the similarities/dissimilarities between cases. The result will be a set of clusters such that cases within a cluster are more similar to each other than they are to cases in other clusters. Cluster analysis has been applied in a wide variety of research problems, from medicine and psychiatry to archeology. In general whenever one needs to classify a large number of information into manageable meaningful piles, or discover similarities between objects, cluster analysis is of great utility.

CLA techniques can be hierarchical (for example the *tree clustering*), i.e. the resultant classification has a increasing number of nested classes, or non-hierarchical when the number of clusters is decided ex ante (for example the *k-means clustering*). Care should be taken that groups (classes) are meaningful in some fashion and are not arbitrary or artificial. To do so the clustering techniques attempt to have more in common with own group than with other groups, through minimization of internal variation while maximizing variation between groups.

Homogeneous and distinct groups are delineated based upon assessment of distances or in the case of Ward's method, an F-test (Davis, 1986). A **distance measure** is an appraisal of the degree of similarity or dissimilarity between cases in the set. A small distance is equivalent to a large similarity. It can be based on a single dimension or on multiple dimensions, for example countries in TAI example can be evaluated according to the TAI composite indicator or they can be evaluated according to all single sub-indicators. Notice that CLA does not "care" whether the distances are real (as in the case of quantitative indicators) or given by the researcher on the basis of an ordinal ranking of alternatives (as in the case of qualitative indicators). Some of the most common distance measures are listed in **Table 3.7** including Euclidean and non-Euclidean distances (e.g. city-block). One problem with Euclidean distances is that they can be greatly influenced by variables that have the largest values. One way around this problem is to standardise the variables.

Table 3.7. *Distance measures* $D(x,y)$ *between two objects* x *and* y *over* N_d *dimensions.*

Euclidean	$D(x,y) = \left(\frac{\sum_{i=1}^{N_d} (x_i - y_i)^2}{N_d} \right)^{1/2}$	This is the geometric distance in a multidimensional space and is usually computed from raw data (prior to any normalization). The advantage is that this measure is not affected by the addition of new objects (for example outliers). Disadvantage: this measure is affected by the difference in scale (e.g. if the same object is measured in centimetres or in meters the *D(x,y)* is highly affected.
Squared Euclidean	$D(x,y) = \frac{\sum_{i=1}^{N_d} (x_i - y_i)^2}{N_d}$	This measure places progressively greater weight on objects that are further apart. Usually this is computed from raw data and shares the same advantages and disadvantages of the Euclidean distance.
City-block (Manhattan)	$D(x,y) = \frac{\sum_{i=1}^{N_d} \lvert x_i - y_i \rvert}{N_d}$	This distance is the average of distances across dimensions and it supplies similar results to the Euclidean distance. In this measure the effect of outliers is less pronounced (since it is not squared). The name comes from the fact that in most American cities it is not possible to go directly between two points, so the route follows the grid of roads.
Chebychev	$D(x,y) = Max \lvert x_i - y_i \rvert$	This measure is mostly used when one wants to define objects as "different" if they are different in any one of the dimensions.

Power	$D(x,y) = \left(\frac{\sum_{i=1}^{N_d} (x_i - y_i)^p}{N_d} \right)^{1/r}$	This distance measure is useful when one wants to increase or decrease the progressive weight that is placed in one dimension, for which the respective objects are very different; *r* and *p* a user-defines parameters: *p* controls the progressive weights placed on differences on individual dimensions, and *r* controls the progressive weight placed on larger differences between objects. For p = r = 2, we have the Euclidean distance.
Percent disagreement	$D(x,y) = \frac{number\ of\ x_i \neq y_i}{N_d}$	Useful if the data are categorical in nature.

Having decided how to measure similarity (the distance measure), the next step is to choose the clustering algorithm, i.e. the rules which govern how distances are measured between clusters. There are many methods available, the criteria used differ and hence different classifications may be obtained for the same data, even using the same distance measure. The most common linkage rules are (Spath, 1980):

- **Single linkage** (nearest neighbor). The distance between two clusters is determined by the distance between the two closest elements in the different clusters. This rule (called also single linkage) produces clusters chained together by single objects.
- **Complete linkage** (farthest neighbor). The distance between two clusters is determined by the greatest distance between any two objects belonging to different clusters. This method usually performs well when objects naturally form distinct groups.
- **Unweighted pair-group average**. The distance between two clusters is calculated as the average distance between all pairs of objects in the two clusters. This method usually performs well when objects naturally form distinct groups. A variation of this method is using the **centroid** of a cluster: the distance is then the average point in the multidimensional space defined by the dimensions.
- **Weighted pair-group average**. Similar to the unweighted pair-group average (centroid included) except for the fact that the size of the cluster (i.e. the number of objects contained) is used as weight for the average distance. This method is useful when cluster sizes are very different.
- **Ward's method** (Ward, 1963). Cluster membership is determined by calculating the variance of elements (the sum of the squared deviations from the mean of the cluster). An element will belong to the cluster is it produces the smallest possible increase in the variance.

Figure 3.2 shows the country clusters based on the technology achievement sub-indicators using tree clustering (hierarchical) with single linkage and squared Euclidean distances. Similarity between countries belonging to the same cluster decreases as the linkage distance increases. One of the biggest problems with CLA is identifying the optimum number of clusters. As the amalgamation process continues increasingly dissimilar clusters must be fused, i.e. the classification becomes increasingly artificial. Deciding upon the optimum number of clusters is largely subjective, although looking at the plot of linkage distance across fusion steps may help (Milligan and Cooper, 1985). Sudden jumps in the level of similarity (abscissa) indicate that dissimilar groups or outliers are fused. Such a plot is presented in Figure 3.3, where the greatest dissimilarity among the 23 countries in the TAI example is found at a linkage distance close to 4.0, which indicates that the data are best represented by ten clusters: Finland alone, Sweden and

USA, the group of countries located between the Netherlands and Hungary, then alone Canada, Singapore, Australia, New Zealand, Korea, Norway, Japan. Notice that the most dissimilar are Korea, Norway and Japan, which are aggregated only at the very end of the analysis. Notice also that this result does not fully correspond to the division in laggard, average and leading countries resulting from the standard aggregation methods. Japan, in fact, would be in the group of leading countries, together with Finland, Sweden, USA, while Hungary, Czech Republic, Slovenia and Italy would be the laggards, far away from the Netherlands, USA or Sweden (see Table 6.11).

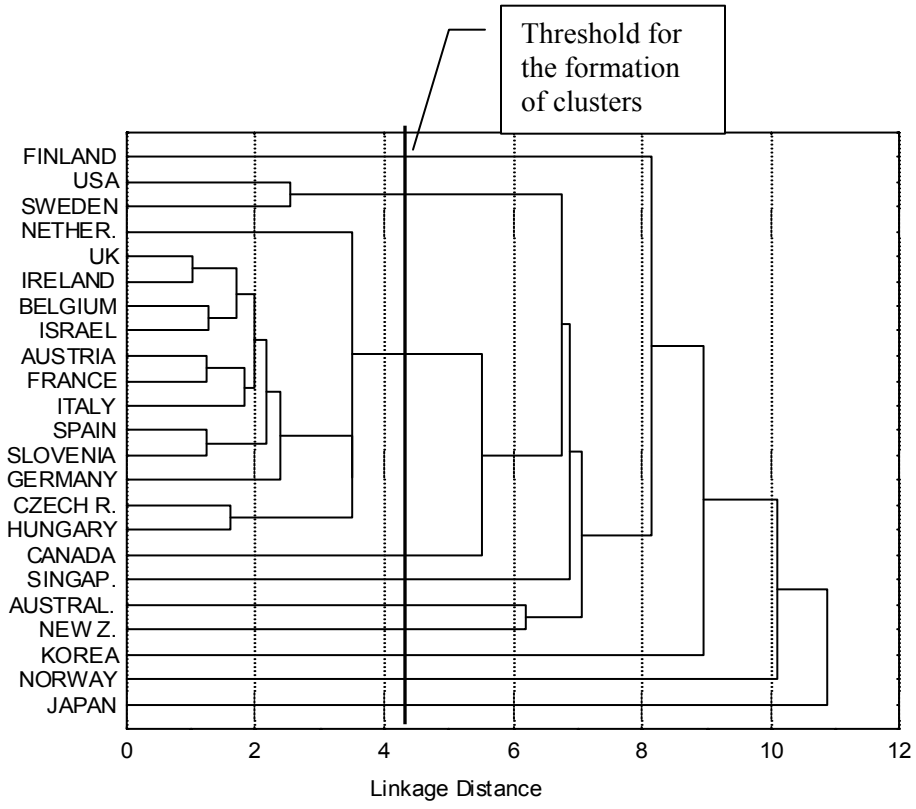


Figure 3.2. *Country clusters for the sub-indicators of technology achievement (standardised data). Type: Hierarchical, single linkage, squared Euclidean distances.*

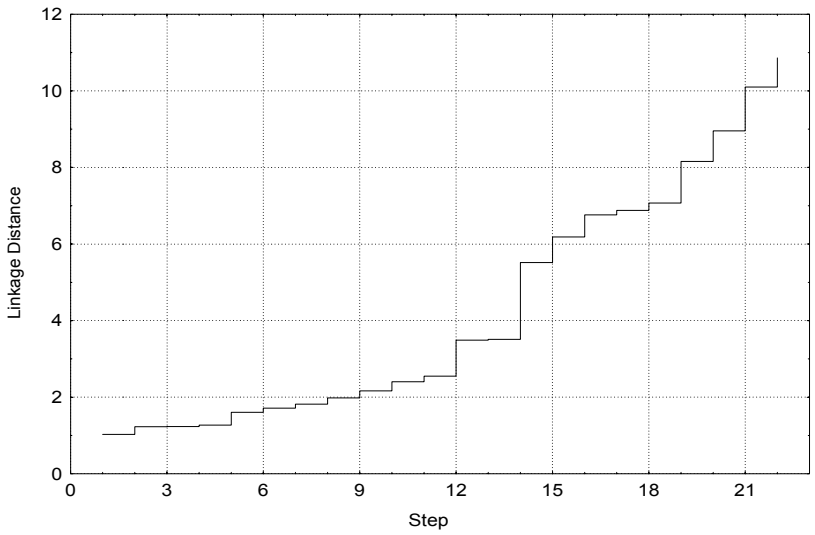


Figure 3.3. *Linkage distance versus fusion step in the hierarchical clustering for the technology achievement example.*

A non-hierarchical method of clustering, different from the Joining (or Tree) clustering shown above, is the **k-means clustering** (Hartigan, 1975). This method is useful when the aim is that of dividing the sample in *k* clusters of greatest possible distinction. The parameter *k* is decided by the analyst, for example we may decide to cluster the 23 countries in the TAI example into 3 groups, e.g. leaders, potential leaders, dynamic adopters. The k-means algorithm will supply 3 clusters as distinct as possible (results shown in Table 3.7). This is done by analyzing the variance of each cluster. Thus, this algorithm can be applied with continuous variables (yet it can be also modified to accommodate for other types of variables). The algorithm starts with *k* random clusters and moves the objects in and out the clusters with the aim of (i) minimizing the variance of elements within the clusters, and (ii) maximize the variance of the elements outside the clusters.

A line graph of the means across clusters is displayed in Figure 3.4. This plot is very useful in summarizing the differences in the means between clusters. It is shown for example that the main difference between the *leaders* and the *potential leaders* (Table 3.9) is on RECEIPTS and EXPORTS. At the same time, the *dynamic adopters* are lagging behind the *potential leaders* due to their lower performance on INTERNET, ELECTRICITY and SCHOOLING. They are, however, performing better on EXPORTS. Two of the sub-indicators, i.e. PATENTS and ENROLMENT, are not useful in distinguishing between these 3 groups, as the cluster means are very close.

Table 3.8. *K-means clustering for the 23 countries in the technology achievement case study*

Group1 (leaders)	Group 2 (potential leaders)	Group 3 (dynamic adopters)	
Finland	Canada	Japan	France
USA	Australia	Korea	Israel
Sweden	Norway	UK	Spain
Netherlands	New Zealand	Singapore	Italy
		Germany	Czech Rep.
		Ireland	Hungary
		Belgium	Slovenia
		Austria	

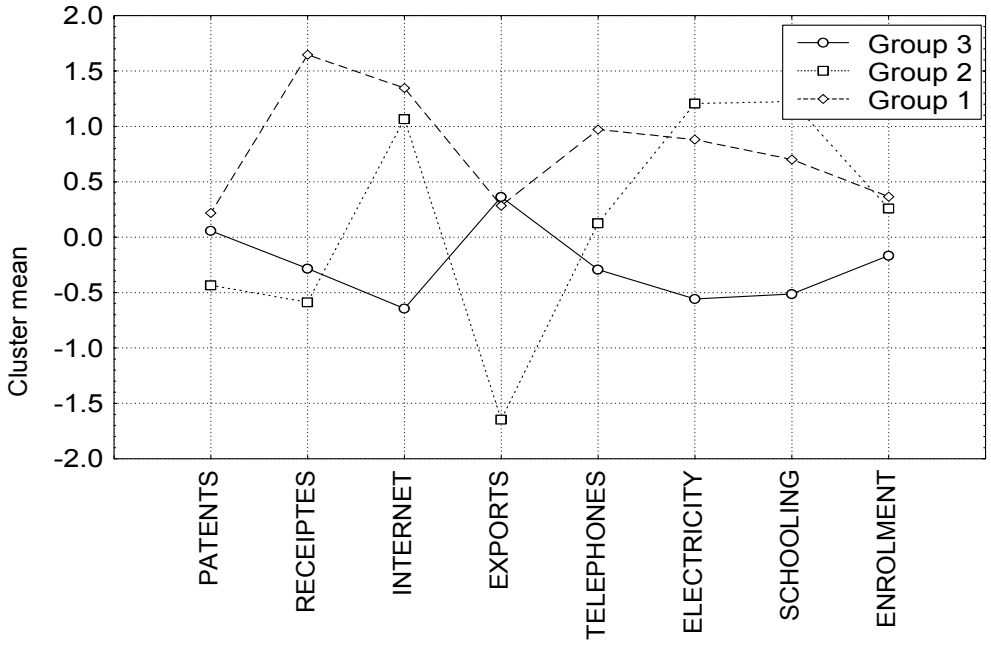


Figure 3.4. *Plot of means for each cluster in the technology achievement case study. Type: k-means clustering (standardized data).*

Finally, expectation maximization (EM) clustering extends the simple k-means clustering in two ways:

1. Instead of clustering the objects by maximizing the differences in means for continuous variables, EM clusters membership on the basis of probability distributions: each observation will belong to each cluster with a certain probability. EM estimates mean and standard deviation of each cluster so as to maximizes the overall likelihood of the data, given the final clusters (Binder, 1981).
2. Unlike k-means, EM can be applied both to continuous and categorical data.

Ordinary significance tests are not valid for testing differences between clusters. This is because clusters are formed to be as much separated as possible, thus the assumptions of usual tests, parametric or non parametric are violated (see Hartigan 1975). As final remark a warning: CLA will always produce a grouping, this means that clusters may or may not prove useful for classifying objects depending upon the objectives of the analysis. For example, if grouping zip code areas into categories based on age, gender, education and income discriminates between wine drinking behaviors, then this would be useful information only if the aim of the CLA was that of establishing a wine store in new areas. Furthermore, CLA methods are not clearly established, there are many options, all giving very different results (see Everitt, 1979).

3.2.2 Factorial k-means analysis

In the previous sections we explored the relationships within a set of variables (e.g. sub-indicators by using continuous models (e.g., Principal Component Analysis or Factor Analysis) that summarize the common information in the data set by detecting non-observable dimensions. On the other hand, the relationships within a set of objects (e.g. countries) are often explored by fitting discrete classification models as partitions, n-trees, hierarchies, via non-parametric techniques of clustering.

When the number of variables is large or when is it believed that some of these do not contribute much to identify the clustering structure in the data set, researchers apply the continuous and discrete models sequentially, frequently carrying out a PCA and then applying a clustering algorithm on the object scores on the first few components. However, De Sarbo et al. (1990), De Soete & Carroll (1994) warn against this approach, called "tandem analysis" by Arabie and Hubert (1994), because PCA or FA may identify dimensions that do not necessarily contribute much to perceive the clustering structure in the data and that, on the contrary, may obscure or mask the taxonomic information.

Various alternative methods combining cluster analysis and the search for a low-dimensional representation have been proposed, and focus on multidimensional scaling or unfolding analysis (e.g., Heiser, 1993, De Soete and Heiser, 1993). A method that combines k-means cluster analysis with aspects of Factor Analysis and PCA is presented by Vichi and Kiers (2001). A discrete clustering model together with a continuous factorial one are fitted simultaneously to two-way data, with the aim to identify the best partition of the objects, described by the best orthogonal linear combinations of the variables (factors) according to the least-squares criterion. This methodology named *factorial k-means analysis* has a very wide range of application since it reaches a double objective: the data reduction and synthesis, simultaneously in direction of objects and variables; Originally applied to short-term macroeconomic data, factorial k-means analysis has a fast alternating least-squares algorithm that extends its application to large data sets. The methodology can therefore be recommended as an alternative to the widely used tandem analysis.

3.3 Conclusions

Application of multivariate statistics, including Factor analysis, Coefficient Cronbach Alpha, Cluster Analysis, is something of an art, and it is certainly not as objective as most statistical methods. Available software packages (e.g. STATISTICA, SAS, SPSS) allow for different variations of these techniques. The different variations of each technique can be expected to give somewhat different results and can therefore confuse the developers of composite indicators. On the other hand, multivariate statistic is widely used to analyse the information inherent in a set of sub-indicators and will continue to be widely used in the future. The reason for this is that developers of composite indicators find the results useful for gaining insight into the structure of their multivariate datasets. Therefore, if it is thought of as a purely descriptive tool, with limitations that are understood, then it must take its place as one of the important steps during the development of composite indicators.

4. Imputation of missing data

Missing data are present in almost all the case studies of composite indicators. Data can be missing either in a *random* or in a *non-random* fashion. They can be missing at random because of malfunctioning equipment, weather issues, lack of personnel, but there is no particular reason to consider that the collected data are substantially different from the data that could not be collected. On the other hand, data are often missing in a non-random fashion. For example, if studying school performance as a function of social interactions in the home, it is reasonable to expect that data from students in particularly types of home environments would be more likely to be missing than data from people in other types of environments. More formally, the missing patterns could be:

- MCAR (Missing Completely At Random): missing values do not depend on the variable of interest or any other observed variable in the data set. For example the missing values in variable *income* would be of MCAR type if (i) people who do not report their income have, on average, the same income as people who do report income, and if (ii) each of the other variables in the dataset would have to be the same, on average, for the people who did not report the income and the people who did report their income.
- MAR (Missing At Random): missing values do not depend on the variable of interest, but they are conditional on some other variables in the data set. For example the missing values in *income* would be MAR if the probability of missing data on income depends on marital status but, within each category of marital status, the probability of missing income is unrelated to the value of income. Missing by design, e.g. if survey question 1 is answered yes, than survey question 2 is not to be answered, are also MAR as missingness depends on the covariates.
- NMAR (Not Missing At Random): missing values depend on the values themselves. For example high income households are less likely to report their income.

One of the problems with missing data is that there is no statistical test for NMAR and often no basis upon which to judge whether data are missing at random or systematically, whilst most of the methods that impute (i.e. fill in) missing values require an MCAR or at least an MAR mechanism. When there are reasons to assume an NMAR pattern, then the missing pattern must be explicitly modelled and included in the analysis. This could be very difficult and could imply ad hoc assumptions that are likely to deeply influence the result of the entire exercise (see Little and Rubin, 2002, chapter 15 for some examples of NMAR mechanisms and Kaufmann, Kraay and Zoid-lobatón, 1999 and 2003 for an application to governance indicators).

Three generic approaches for dealing with missing data can be distinguished, i.e. case deletion, single imputation or multiple imputation. The first one, *Case Deletion,* simply omits the missing records from the analysis. The disadvantages of this approach (also called complete case analysis) are that it ignores possible systematic differences between complete and in-complete sample and produces unbiased estimates only if deleted records are a random sub-sample of the original sample (MCAR assumption). Furthermore, standard errors will, in general be larger in a reduced sample given that less information is used. As a rule of thumb (Little and Rubin, 1987) if a variable has more than 5% missing values, cases are not deleted, and many researchers are much more stringent than this.

The other two approaches see the missing data as part of the analysis and therefore try to impute values through either *Single Imputation* (e.g. Mean/Median/Mode substitution, Regression Imputation, Expectation-Maximisation Imputation, etc.) or *Multiple Imputation* (e.g. Markov Chain Monte Carlo algorithm). The advantages of imputation include the minimisation of bias

and the use of 'expensive to collect' data that would otherwise be discarded. The main disadvantage of imputation is that it can allow data to influence the type of imputation. In the words of Dempster and Rubin (1983):

> The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can legitimately handled in this way and situations where standard estimators applied to real and imputed data have substantial bias.

The uncertainty in the imputed data should be reflected by variance estimates. This allows taking into account the effects of imputation in the course of the analysis. However, Single Imputation is known to underestimate the variance, because it reflects partially the imputation uncertainty. The Multiple Imputation method instead, which provides several values for each missing value, can more effectively represent the uncertainty due to imputation. No imputation model is free of assumptions and the imputation results should hence be thoroughly checked for their statistical properties such as distributional characteristics as well as heuristically for their meaningfulness, e.g. whenever negative imputed values are possible.

This section illustrates the main issues related to imputation. The literature on the analysis of missing data is extensive and in rapid development. Therefore, this section is not intended to be comprehensive, but rather to supply the reader with the basic flavour of the main methods. More comprehensive surveys can be found in Little and Rubin (2002), Little (1997) and Little and Schenker (1994).

4.1 Single imputation

As indicated by Little and Rubin (2002), imputations are means or draws from a predictive distribution of the missing values. The predictive distribution must be created by employing the observed data. There are, in general, two approaches to generate this predictive distribution:

Implicit modeling: the focus is on an algorithm, with implicit underlying assumptions that should be assessed. Besides the need to carefully verify whether the implicit assumptions are reasonable and fit to the issue dealt with, the danger of this type of modeling missing data is to consider the resulting data set as complete and forget that an imputation has been done. Implicit modeling includes:

- **Hot deck imputation**: fill in blanks cells with individual data drawn from "similar" responding units, e.g. missing values for individual income may be replaced with the income of another respondent with similar characteristics (age, sex, race, place of residence, family relationships, job, etc.).
- **Substitution**: replace non responding units with units not selected into the sample, e.g. if a household cannot be contacted, then a previously non selected household in the same housing block is selected.
- **Cold deck imputation**: replace the missing value with a constant value from an external source, e.g. from a previous realization of the same survey.

Explicit modeling: the predictive distribution is based on a formal statistical model where the assumptions are made explicit. This is the case of the

- **Unconditional mean/median/mode imputation**, where the sample mean (median, mode) of the recorded values for the given sub-indicator substitutes the missing values.
- **Regression imputation**. Missing values are substituted by the predicted values obtained from a regression. The dependent variable of the regression is the sub-indicator hosting the missing value and the regressor(s) is(are) the sub-indicator(s) showing a strong relationship with the dependent variable (usually a high degree of correlation).
- **Expectation Maximization (EM) imputation**. This model focuses on the interdependence between model parameters and the missing values. The missing values are substituted by estimates obtained through an iterative process. First, one predicts the missing values based on initial estimates of the model parameter values. These predictions are then used to update the parameter values, and the process is repeated. The sequence of parameters converges to the maximum likelihood estimates, and the time to converge depends on the proportion of missing data and the flatness of the likelihood function.

If the simplicity is its main appeal, an important limitation of the single imputation methods is that they systematically underestimate the variance of the estimates (with some exceptions for the EM method where the bias depends on the algorithm used to estimate the variance). Therefore, they do not fully allow assessing the implications of imputation and thus the robustness of the composite index derived from the imputed dataset.

3.1.1 Unconditional mean imputation

Let X_q be the random variable associated to the sub-indicators $q=1,...,Q$ and $x_{q,c}$ the observed value of X_q for country c, with $c=1,..,M$. For some c indicate with m_q the number of recorded values on X_q, and $M-m_q$ the number of missing values. The unconditional mean will be calculated as

$$\bar{x}_q = \frac{1}{m_q} \sum_{recorded} x_{q,c} \qquad (4.1)$$

Similarly, the median (the value that divides in two equal parts the distribution of the random variable) and the mode (the value with the highest frequency) of the distribution would be calculated on the available sample and substitute missing values.[4] The consequences of "fill in" blank spaces with the sample mean is that the imputed value is a biased estimator of the population mean (except in the case of MCAR mechanisms) and the sample variance underestimates true variance with the consequence of underestimating the uncertainty on the composite due to the imputation.

[4] A variant of unconditional mean imputation is the fill-in via conditional mean. The regression approach is one possible method. Another common method (called imputing means within adjustment cells) is to classify the data for the sub-indicator with some missing values in classes and impute provisionally the missing values of that class with the sample mean of the class. Then sample mean (across al classes) is then calculated and substituted as final imputation value.

4.1.2 Regression imputation

Suppose to have a set of $h\text{-}1<Q$ fully observed sub-indicators ($x_1,\ldots,x_{h-1}$) and a sub-indicator x_h only observed for r countries but missing for the remaining M-r countries. Regression imputation computes the regression of x_h on ($x_1,\ldots,x_{h-1}$) using r complete observations, and impute the missing values as prediction from the regression[5]:

$$\hat{x}_{ih} = \hat{\beta}_0 + \sum_{j=1}^{h-1} \hat{\beta}_j x_{ij} \quad i = 1,..,M-r \tag{4.2}$$

Usually the strategy to define the 'best' regression is a two step procedure. First, all different subsets of predictors are adopted in a multiple regression manner. Then, the best subset(s) is determined using the following criteria:[6]

- the value of R^2
- the value of the residual mean square *RMS*
- the value of Mallows' C_k
- stepwise regression

A variation of the regression approach is the stochastic regression approach that imputes a conditional draw instead of imputing the conditional mean:

$$\hat{x}_{ih} = \hat{\beta}_0 + \sum_{j=1}^{h-1} \hat{\beta}_j x_{ij} + \varepsilon_i \quad i = 1,..,M-r \tag{4.2}$$

where ε_i is a random variable $N(0,\hat{\sigma}^2)$ and $\hat{\sigma}^2$ is the residual variance from the regression of x_h on ($x_1,\ldots,x_{h-1}$) based on the r complete cases.

A key problem of both approaches is again the underestimation of the standard errors (although stochastic regression ameliorates the distortions), thus the inference based on the entire dataset (including the imputed data) does not fully count for imputation uncertainty. The result is that *p-values* of tests are too small and confidence intervals too narrow. Replication methods and multiple imputation are likely to correct the loss of precision of simple imputation.

What if the variable with missing information is categorical? Regression imputation is still possible but adjustments using, e.g. rounding of the predictions or a logistic, ordinal or multinomial logistic regression models, are required. For nominal variables, frequency statistics such as the mode or hot- and cold-deck imputation methods might be more appropriate.

4.1.3 Expected maximization imputation

Suppose that X denotes the data. In the likelihood based estimation the data are assumed to be generated by a model described by a probability or density function $f(X/\theta)$, where θ is the unknown vector parameter vector lying in the parameter space Ω_θ (e.g. the real line for means, the positive real line for variances and the interval [0,1] for probabilities). The probability function captures the relationship between the data set and the parameter of the of the data model

---

[5] If the observed variables are dummies for a categorical variable then the prediction (4.2) are respondent means within classes defined by the variable and the method reduces to that of imputing means with adjustment cells.

[6] Define $SSE = \sum_i (x_{ih} - \hat{x}_{ih})^2$, $SST = \sum_i (x_{ih} - \bar{x}_h)^2$, then $R^2 = 1 - (SSE/SST)$, $MSE = SSE/(M-r-k)$, where k is the number of coefficients in the regression and *(M-r)* the number of observations. $RMS = \sum_i (\hat{x}_{ih} - \bar{x}_h)^2$ and $C_k = (SSE_k / MSE) - (M-r) + 2k$ where the SSE_k is computed from a model with only k coefficients and MSE is computed using all available regressors.

and describes the probability of observing a dataset for a given $\theta \in \Omega_\theta$. Since θ is unknown while the data set is known, it make sense to reverse the argument and look for the probability of observing a certain θ given the data set *X*: this is the likelihood function. Therefore, given *X*, the likelihood function $L(\theta / X)$ is any function of $\theta \in \Omega_\theta$ proportional to $f(X / \theta)$:

$$L(\theta / X) = k(X) f(X / \theta) \tag{4.3}$$

Where $k(X) > 0$ is a function of *X* and not of θ. The log-likelihood is then the natural logarithm of the likelihood function. In the case of *M* independent and identically distributed observations $X = (x_1, ..., x_M)^T$, from a normal population with mean μ and variance σ^2 the joint density is

$$f(X / \mu, \sigma^2) = (2\pi\sigma^2)^{-M/2} \exp\left(-\frac{1}{2}\sum_{c=1}^{M}\frac{(x_c - \mu)^2}{\sigma^2}\right) \tag{4.4}$$

For a given sample *X* the log-likelihood is (ignoring additive constants of function $f(\cdot)$) a function of (μ, σ^2):

$$\begin{aligned} l(\mu, \sigma^2 / X) &= ln[L(\mu, \sigma^2 / X)] = ln[k(X) f(X / \mu, \sigma^2)] \\ &= ln\, k(X) - \frac{M}{2} ln\, \sigma^2 - \frac{1}{2}\sum_{c=1}^{M}\frac{(x_c - \mu)^2}{\sigma^2} \end{aligned} \tag{4.5}$$

Maximizing the likelihood function corresponds to the question of which value of $\theta \in \Omega_\theta$ is mostly supported by a given sampling realization *X*. This implies solving the likelihood equation:

$$D_l(\theta / X_{obs}) \equiv \frac{\partial ln\, L(\theta / X_{obs})}{\partial \theta} = 0 \tag{4.6}$$

When a closed-form solution of equation (4.6) cannot be found, iterative methods can be applied. The EM algorithm is one of these iterative methods. [7] The issue is that *X* contains both observable and missing values, i.e. $X = (X_{obs}, X_{mis})$. Thus one has to find both the unknown parameters and the unknown observations of the model.

Assume that missing data are MAR or MCAR[8], the EM consists of two components, the expectation (E) and maximization (M) steps. Each step is completed once within each algorithm cycle. Cycles are repeated until a suitable convergence criterion is satisfied. In the M step the maximum likelihood estimation of θ is computed just as if there were no missing data (thus missing values are replaced by estimated values, i.e. initial conditions in the first round of maximization). In the E step the missing data are estimated by their expectations given the observed data and current estimated parameter values. In the following maximization step the

[7] Other iterative methods include the Newton-Raphson algorithm and the scoring method. Both involve a calculation of the matrix of second derivatives of the likelihood, which, for complex pattern of incomplete data, can be a very complicate function of θ. As a result these algorithms often require algebraic manipulations and complex programming. Numerical estimation of this matrix is also possible but careful computation is needed.

[8] For NMAR mechanisms one needs to make assumption on the missing-data mechanism and include them into the model, see Little and Rubin, 2002, Ch. 15.

parameters in θ are re-estimated using maximum likelihood applied to the observed data augmented by the estimates of the unobserved data (coming from the previous round). The whole procedure is iterated until convergence (absence of changes in estimates and in the variance-covariance matrix). Effectively, this process maximizes, in each cycle, the expectation of the complete data log likelihood. On convergence, the fitted parameters are equal to a local maximum of the likelihood function (which is the maximum likelihood in the case of a unique maximum).

The advantage of the EM is its broadness (it can be used for a broad range of problems, e.g. variance component estimation or factor analysis), its simplicity (EM algorithm are often easy to construct conceptually and practically), and each step has a statistical interpretation and convergence is reliable. The main drawback is that in some cases, with a large fraction of missing information, convergence may be very slow. The user should also care that the maximum found is indeed a global maximum and not a local one. To test this, different initial starting values for each θ can be used.

4.2 Multiple imputation

Multiple imputation (MI) is a general approach that does not require a specification of paramentrized likelihood for all data. The idea of MI is depicted in Figure 4.1. The imputation of missing data is performed with a random process that reflects uncertainty. Imputation is done N times, to create N "complete" datasets. On each dataset the parameter of interest are estimated, together with their standard errors. Average (mean or median) estimates are combined using the N sets and between and within imputation variance is calculated.

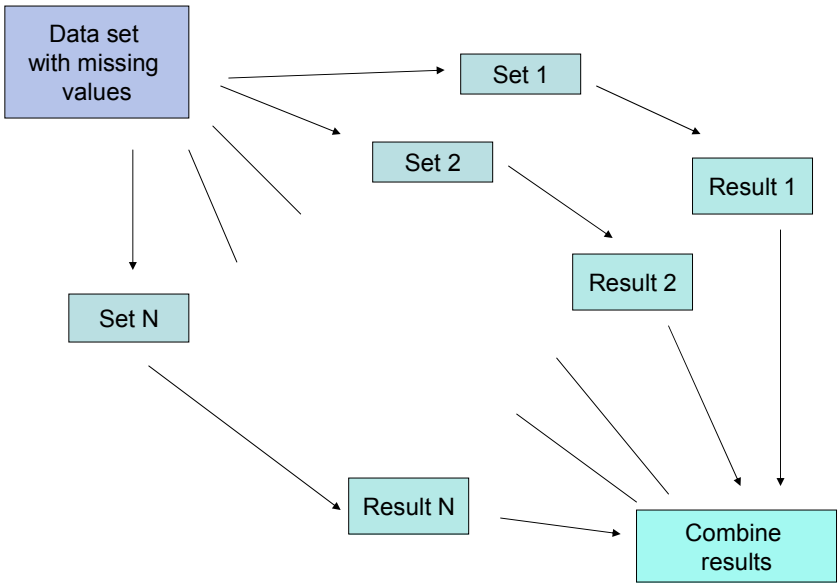


Figure 4.1. *Logic of multiple imputation*

Any "proper" imputation method can be used in multiple imputation. For example, one could use regression imputation repeatedly, drawing *N* values of the regression parameters using the variance matrix of estimated coefficients. However, one of the most general models is the Markov Chain Monte Carlo (MCMC) method. Markov chain is a sequence of random variables

in which the distribution of the actual element depends on the value of the previous one. It assumes that data are drawn from a multivariate Normal distribution and requires MAR or MCAR assumptions.

The theory of MCMC is most easily understood using Bayesian methodology (See Figure 4.2). Let us denote the observed data as X_{obs} and the complete dataset as $X=(X_{obs}, X_{mis})$, where X_{mis} is to be filled in via multiple imputation. If the distribution of X_{mis}, with parameter vector θ, were known then we could impute X_{mis} by drawing from the conditional distribution $f(X_{mis}|X_{obs}, \theta)$[9]. However, since θ is unknown, we shall estimate it from the data, yielding $\hat{\theta}$, and use the distribution $f(X_{mis}|X_{obs}\hat{\theta})$. Since $\hat{\theta}$ is itself a random variable, we must also take its variability into account in drawing imputations.

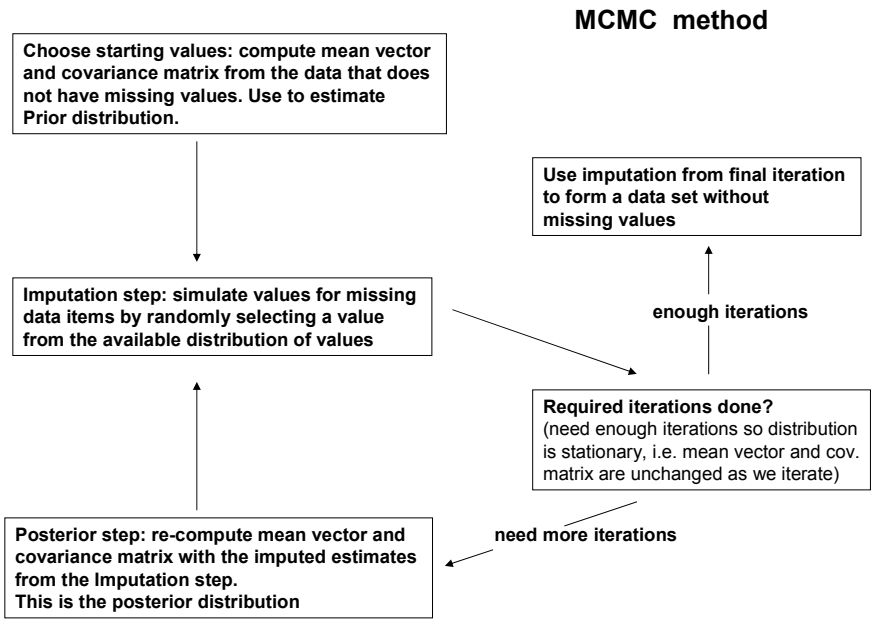


Figure 4.2. *Functioning of MCMC method*[10]

In Bayesian terms, θ is a random variable whose the distribution depends on the data. So the first step for its estimation is to obtain the posterior distribution of θ from the data,.Usually this posterior is approximated by a normal distribution. After formulating the posterior distribution of θ, the following imputation algorithm can be used:

- Draw θ^* from the posterior distribution of θ, $f(\theta|Y, X_{obs})$ where Y denotes exogenous variables that may influence θ.
- Draw X_{mis} from $f(X_{mis}|Y, X_{obs}, \theta^*)$

[9] The missing data generating process may depend on additional parameters φ, but if φ and θ are independent, the process called ignorable and the analyst can concentrate on modelling the missing data, given the observed data and θ. If the two processes are not independent, then we have non-ignorable missing data generating process, which cannot be solved adequately without making assumptions on the functional form of the interdependency.

[10] rearranged from K. Chantala and C. Suchindran, http://www.cpc.unc.edu/services/computer/presentations/mi_presentation2.pdf

- Use the completed data X and the model to estimate the parameter of interest (e.g. the mean) β^* and its variance V(β^*) (within-imputation variance).

These steps are repeated independently N times, resulting in β_n^* , $V(\beta_n^*)$, $n=1,...,N$. Finally, the N imputations are combined. A possible *combination* is the mean of all individual estimates (but also the median can be used):

$$\beta^* = \frac{1}{N}\sum_{n=1}^{N}\beta_n^* \tag{4.7}$$

This *combination* will be the value that fills in the blank space in the dataset. The total variance is obtained as a weighted sum of the *within-imputation* variance and the *between-imputations* variance:

$$V^* = \overline{V} + \frac{N+1}{N}B \tag{4.8}$$

where the mean of the *within-imputation* variances is

$$\overline{V} = \frac{1}{N}\sum_{n=1}^{N}V\left(\beta_n^*\right) \tag{4.9}$$

and the *between-imputations* variance is given by

$$B = \frac{1}{N-1}\sum_{n=1}^{N}\left(\beta_n^* - \beta^*\right)\left(\beta_n^* - \beta^*\right)' \tag{4.10}$$

Confidence intervals are obtained by taking the overall estimate plus or minus a multiple of standard error, where that number is a quantile of Student's t-distribution with degrees of freedom:

$$df = (N-1)\left(1+\frac{1}{r}\right)^2 \tag{4.11}$$

where r is the *between-to-within* ratio.

$$r = \left(1+\frac{1}{N}\right)\frac{B}{\overline{V}} \tag{4.12}$$

Based on these variances, one can calculate approximate 95% confidence intervals.

In conclusion, *Multiple Imputation* method imputes several values (N) for each missing value (from the predictive distribution of the missing data), to represent the uncertainty about which values to impute. The N versions of completed data sets are analyzed by standard complete data methods and the results are combined using simple rules to yield single combined estimates (e.g., MSE, regression coefficients), standard errors, p-values, that formally incorporate missing data uncertainty. The pooling of the results of the analyses performed on the multiply imputed data sets, implies that the resulting point estimates are averaged over the N completed sample points, and the resulting standard errors and p-values are adjusted according to the variance of the corresponding N completed sample point estimates. Thus, the '*between imputation variance*',

provides a measure of the extra inferential uncertainty due to missing data (which is not reflected in single imputation).

5. Normalisation of data

The indicators selected for aggregation convey at this stage quantitative information of different kinds[11]. Some indicators can be incommensurate with others, and have different measurement units. In the TAI, for example, the number of patents granted to residents is expressed per capita, and the high and medium technology exports are expressed as percentage of total exports.

Therefore, to avoid adding up apples and pears, before going to the aggregation stage it is necessary to bring the indicators to the same standard, by transforming them in pure, dimensionless, numbers. We call this process *normalization*. There are a number of such methods available: the most commonly encountered in the literature are reviewed in this section.

The objective is to identify the most suitable normalization procedures to apply to the problem at hand, taking into account their properties with respect to the measurement units in which the indicators are expressed, and their robustness to possible outliers in the data. Different normalization methods will supply different results for the composite indicator. Therefore, overall robustness tests should be carried out to assess their impact on the outcomes.

5.1 Scale transformations

There is an aspect which the normalization process may interfere with. This is the scale effect, i.e. the different measurement units in which an indicator can be expressed before its normalization. Some normalization procedures are invariant to changes in measurement unit of the indicator, as they provide the same normalized values whatever the measurement unit of the indicator is. That is, temperature could be expressed equivalently in Celsius or Fahrenheit and the result of the normalization is not affected. Other normalizations unfortunately are not invariant. Applying a normalization procedure which is not invariant to changes in the measurement unit could result in different outcomes for the composite indicator. Let us give a very simple example with two indicators (temperature and humidity) for two countries A and B for two different years. The raw data are given in Table 5.1, where we assume that the temperature is expressed in Celsius.

Table 5.1 *Raw data on temperature (in Celsius) and humidity for two countries A and B*

	2003	2004
Country A –Temperature (ºC)	35	35.9
Country A –Humidity (%)	75	70
Country B –Temperature (ºC)	39	40
Country B –Humidity (%)	50	45

We normalize each indicator by dividing by the value possessed by the country leader, and then aggregate the two indicators by applying equal weights. The result is given in table 5.2, where we can see that Country A increases its performance with time.

[11] This chapter is based on the state-of-the-art report (JRC, 2002), the report from OECD (Freudenberg, 2003) and the technical paper by Jacobs et al., (2004).

Table 5.2 *Composite indicator for A and B obtained with normalization based on "distance to the best performer" and temperature expressed in Celsius.*

	2003	2004
Country A	0.94872	0.94875
Country B	0.83333	0.82143

Now, assume that the same temperature is expressed in Fahrenheit (see Table 5.3).

Table 5.3 *Raw data on temperature (in Fahrenheit) and humidity for two hypothetical countries A and B*

	2003	2004
Country A –Temperature (F)	95	96.62
Country A –Humidity (%)	75	70
Country B –Temperature (F)	102.2	104
Country B –Humidity (%)	50	45

Using the same normalization procedure and aggregating by equal weights, the result shows a completely different pattern, i.e. that the composite indicator for country A now decreases with time (see Table 5.4).

Table 5.4 *Composite indicator for A and B obtained with normalization based on "distance to the best performer" and temperature expressed in Fahrenheit*

	2003	2004
Country A	0.964775	0.964519
Country B	0.83333	0.82143

This means that, when applying this specific normalization procedure, the measurement unit in which the indicator is expressed influences the outcome of the analysis.
On the other hand, using the method of standardization, described in this section, which is invariant to changes in measurement unit, we obtain exactly the same values for the composite indicator, whatever the unit of measurement for the temperature is.

The example illustrated so far is a case of 'interval scale', based on a transformation *f* defined as:
$f: \quad x \to y = \alpha x + \beta; \alpha > 0, \beta \neq 0$.
Here the variable *x* is the temperature expressed in Celsius and y is the temperature expressed in Fahrenheit. Their relationship is indeed:

$$y(\mathrm{F}) = \frac{9}{5} x(^{\circ}\mathrm{C}) + 32$$

Another common change of measurement unit is the so-called 'ratio scale', which is based on the transformation:
$f: \quad x \to y = \alpha x; \alpha > 0$.
To give an example, a "length" might be expressed in centimeters (cm) or yards (yd). Their relationship is indeed: 1 yd = 91.44 cm. The normalization by country leader, not invariant on the 'interval scale', is invariant on the 'ratio scale'. In general, all normalizations that are invariant on the 'interval scale', are also invariant on the 'ratio scale'.

Another transformation which is often used to reduce the skewness of (positive) data varying across many orders of magnitudes is the logarithmic transformation:
$f: \quad x \to y = \log(x); x > 0$.

More in detail, when the range of country-based values for the indicator at hand is wide, or it is positively skewed, the log transformation shrinks the range on its right-hand side. As values approach zero they are also penalised because, after transformation, they become largely negative. When the weighted variables in a linear aggregation are expressed in logarithms, this is equivalent to the geometric aggregation of the variables without logarithms. The ratio between two weights indicates the percentage improvement in one indicator that would compensate for a one percentage point decline in another indicator. This transformation leads to attributing higher weight for a one unit improvement starting from a low level of performance, compared to an identical improvement starting from a high level of performance.

The normalization methods described in this section are all non invariant to this type of scale transformation. The user may decide to use or not the log transformation before the normalization, yet s/he has to beware that the normalized data will surely be affected by the log transformation.

A note on outliers[12] before starting the description of the most commonly used normalisation approaches. Outliers can, in some circumstances, reflect the presence of unwanted information. Therefore, data have to be processed via specific treatment. An example is offered in the Environmental Sustainability Index, where the variable distributions outside the 2.5 and 97.5 percentile scores are trimmed to partially correct for outliers as well as to avoid having extreme values overly dominate the aggregation algorithm. Any observed value greater than the 97.5 percentile is lowered to equal the 97.5 percentile. Any observed value lower than the 2.5 percentile is raised to equal the 2.5 percentile. It is advisable to first try to remove outliers, and consequently perform the normalisation, as this latter procedure can be more or less sensitive to outliers.

5.2 Normalisation methods

5.2.1 Ranking of indicators across countries

The simplest normalisation method consists in ranking each indicator across countries. The main advantages of this approach are its simplicity and the independence to outliers. Disadvantages are the loss of information on absolute levels and the impossibility to draw any conclusion about difference in performance.

This method has been employed to build a composite on the development and application of information and communication technology across countries (see Fagerberg, 2001) and also in the Medicare study on healthcare performance across US States (Jencks et al., 2003).

[12] In a sample of *n* observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are not from a different population, or that the sampling technique is a fault. Such values are called outliers (F.H.C. Marriott, 1990, A dictionary of statistical terms, Longman Scientific & Technical, Fifth edition, p.223). Eurostat adopts this definition of outlier.

For time-dependent studies, the ranking is carried out at each point in time. Therefore we can follow country performance in terms of relative positions (rankings). However, this does not allow the user to follow the absolute performance of each country across time: perhaps the country improves from one year to the next, yet its ranking deteriorates as other countries improve faster.

5.2.2 Standardisation (or z-scores)

For each sub-indicator x_{qc}^t, the average across countries $x_{qc=\bar{c}}^t$ and the standard deviation across countries $\sigma_{qc=\bar{c}}^t$ are calculated. The normalization formula is: $I_{qc}^t = \frac{x_{qc}^t - x_{qc=\bar{c}}^t}{\sigma_{qc=\bar{c}}^t}$, so that all the I_{qc}^t have similar dispersion across countries. The actual minima and maxima of the I_{qc}^t across countries depend on the sub-indicator.

It is the most commonly used because it converts all indicators to a common scale with an average of zero and standard deviation of one. The average of zero means that it avoids introducing aggregation distortions stemming from differences in indicators means. The scaling factor is the standard deviation of the indicator across the countries. Thus, an indicator with extreme values will have intrinsically a greater effect on the composite indicator. This might be desirable if the intention is to reward exceptional behaviour, that is, if an extremely good result on few indicators is thought to be better than a lot of average scores. This effect can be corrected in the aggregation methodology, e.g. by excluding the best and worst sub-indicator scores from the inclusion in the index or by assigning differential weights based on the "desirability" of the sub-indicators scores.

This method is used for the two composite indicators of the knowledge-based economy, published by the European Commission on Key Figures 2003-2004, for the environmental sustainability index developed at Yale University, and in the internal market index 2002. Also, the WHO index of health system performance has been criticized for not using appropriate method of transformation and the z-scores transformation has been recommended (SPRG, 2001).

For time – dependent studies, in order to assess country performance across years, the average across countries $x_{qc=\bar{c}}^{t_0}$ and the standard deviation across countries $\sigma_{qc=\bar{c}}^{t_0}$ are calculated for a reference year (usually the initial time point t_0).

5.2.3 Re-scaling

Each indicator x_{qc}^t for a generic country c and time t is transformed in $I_{qc}^t = \frac{x_{qc}^t - min_c(x_q^t)}{max_c(x_q^t) - min_c(x_q^t)}$ where $min_c(x_q^t)$ and $max_c(x_q^t)$ are the minimum and the maximum value of x_{qc}^t across all the countries *c* at time t. In this way, the normalized indicators I_{qc} have values laying between 0 (laggard, $x_{qc}^t = min_c(x_q^t)$), and 1 (leader, $x_{qc}^t = max_c(x_q^t)$.

Here the transformation is based on the range rather than on the standard deviation. This procedure normalizes the indicators so that they all have identical range (0 1). The extreme values (minimum and maximum) could be unreliable outliers, and have a distortion effect on the transformed indicator. On the opposite, for indicator values lying within an interval with very small range, this latter is widened applying the re-scaling, thus explicitly increasing the effect on the composite indicator (more than they would using the z-scores transformation).

The expression $I_{qc}^{t} = \frac{x_{qc}^{t} - min_c(x_q^{t_0})}{max_c(x_q^{t_0}) - min_c(x_q^{t_0})}$ is sometimes used for time-dependent studies. However, because the drawback is that, if $x_{qc}^{t} > max_c(x_q^{t_0})$, the normalised indicator y_{qc}^{t} would be larger than 1.

Another variant of the rescaling method is the one taking into account the evolution of indicators across time: $I_{qc}^{t} = \frac{x_{qc}^{t} - min_{t\in T}\, min_c(x_q^{t})}{max_{t\in T}\, max_c(x_q^{t}) - min_{t\in T}\, min_c(x_q^{t})}$ where we calculate minimum and maximum for each indicator both across countries and across the whole time range T of the analysis. With this formula the normalized indicators I_{qc}^{t} have values between 0 and 1. However, this transformation is not stable when data for a new time point become available. This implies an adjustment of the analysis period T, which may, in turn, affect the minimum and the maximum for some sub-indicators and, therefore, the values of the I_{qc}^{t} themselves. In such cases, to maintain comparability between the existing and the new data, the composite indicator would have to be recalculated for the existing data.

5.2.4 Distance to a reference country

This method takes the ratios of the indicator x_{qc}^{t} for a generic country c and time t with respect to the sub-indicator $x_{qc=\bar{c}}^{t_0}$ for the reference country at the initial time t_0.

$$I_{qc}^{t} = \frac{x_{qc}^{t}}{x_{qc=\bar{c}}^{t_0}}$$

Using the denominator $x_{qc=\bar{c}}^{t_0}$, the transformation takes into account the evolution of indicators across time; alternatively one can use the denominator $x_{qc=\bar{c}}^{t}$, with running time t.

The reference could be a target to be reached in a given time frame. For example, in the Kyoto protocol, 8% reduction target is established for CO_2 emissions within 2010 for the EU members. This approach is used in the Environmental Policy Performance Indicator (Adriaanse, 1993). The study aims to monitor the trend in the total environmental pressure in the Netherlands and to indicate whether environmental policies are heading in the right direction. The reference could also be an external benchmarking country. For example, United States or Japan are benchmark countries for the composite indicators built in the frame of the EU Lisbon agenda. The reference country could alternatively be the average country within the group of countries considered in the analysis. Here, the average country will be given value 1, and the countries receive scores depending on their distance from the average country. Indicators that are higher than 1 after transformation show countries with above-average performance. The reference country could also be the group leader (‘distance from the best performer’). The value 1 is given to the leading

country and the others are given percentage points away from the leader. The disadvantage is that this approach is based on extreme values which could be unreliable outliers.

A different approach is to consider the country itself as the reference country and calculate the distance in terms of the initial time point as

$$I_{qc}^{t} = \frac{x_{qc}^{t}}{x_{qc}^{t_0}}.$$

This approach is used in *Concern about environmental problems* (Parker, 1993) for measuring the concern of the public on certain environmental problems in three countries (Italy, France and the UK) and in the European Union.

Another kind of distance can be used for the normalisation:

$$y_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc=\bar{c}}^{t_0}}{x_{q=\bar{c}}^{t_0}}$$

This is essentially equal to the one above: instead of being centred on one, it is centred on zero. In the same way, the reference country can either be the average country, or the group leader, or, finally, an external benchmark.

5.2.5 Categorical scales

Each indicator is assigned a categorical score. First, the categories are selected. They can be numerical, such as one, two or three stars, or qualitative, such as 'fully achieved', 'partly achieved' or 'not achieved'. Each category is then assigned a score, which is, to a certain extent, arbitrary.

Often, the scores are based on the percentiles of the distribution of the indicator across the countries. For example, the top 5% of the units receive a score of 100, the units between the 85th and 95th percentiles receive 80 points, the units between the 65th and the 85th percentiles receive 60 points, the units between the 35th and the 65th percentiles receive 50 points, the units between the 15th and the 35th percentiles receive 40 points, the units between the 5th and the 15th percentiles receive 20 points, and, finally, the bottom 5% of the units receive 0 points (see Table 1). This is a way to prize the most performing countries and penalize the less performing ones.

An advantage of this transformation is that, using the same percentile transformation for different years, any small change in the definition of the indicator that could occur with time will not affect the transformed variable. However, in this way we will not be able to track improvements year by year.

Categorical scales omit a large amount of information about the variance between units in the transformed indicators. Another disadvantage is that, if there is little variation within the original scores, the percentile banding forces the categorization on the data, irrespective of the distribution of the underlying data. One possible solution to this is to adjust the percentile brackets across the individual indicators in order to obtain transformed categorical variables with almost normal distributions.

This type of normalization can be found in Nicoletti et al. (2003), an OECD report describing the construction of summary indicators from a large OECD database of economic and administrative product market regulations and employment protection legislation. The summary indicators help to compare the economic and administrative regulatory environment across countries. The summary indicators are obtained by means of factor analysis, in which each component of the regulatory framework is weighted according to its contribution to the overall variance in the data. Data have been gathered basically from Member countries responses to the OECD Regulatory Indicators Questionnaire, which include both qualitative and quantitative information. Qualitative information is coded by assigning a numerical value to each of its possible modalities (e.g. ranging from a negative to an affirmative answer) while the quantitative information (such as data on ownership shares or notice periods for individual dismissals) is subdivided into classes. Then, the resulting coded information is normalised by ranking it on a common 0-6 scale, reflecting the increasing restrictiveness of the regulatory provisions.

5.2.6 Indicators above or below the mean

This transformation considers the indicators that are above, and below, an arbitrarily defined threshold *p* around the mean. The formula employed is:

$$\textbf{if } x_{qc}^{t} \Big/ x_{qc=\bar{c}}^{t_0} > (1+p) \textbf{ then } I_{qc}^{t} = 1$$

$$\textbf{if } x_{qc}^{t} \Big/ x_{qc=\bar{c}}^{t_0} < (1-p) \textbf{ then } I_{qc}^{t} = -1$$

$$\textbf{if } (1-p) < x_{qc}^{t} \Big/ x_{qc=\bar{c}}^{t_0} < (1+p) \textbf{ then } I_{qc}^{t} = 0$$

The threshold builds a neutral region around the mean, where the transformed indicator is zero. This aims at reducing the sharp discontinuity (from -1 to +1) that would exist across the mean value, to two minor discontinuities (from -1 to 0 and from 0 to +1) that exist across the thresholds. A larger number of thresholds could be created at different distances from the mean value. However, this method would overlap with the transformation based on categorical scales.
The advantage of this transformation is its simplicity and the fact that is not affected by outliers. The disadvantages are the arbitrariness of the threshold level and the omission of absolute level information. For example, assume that the value of a given indicator for country A is 3 times (300%) above the mean calculated across all the countries, and the value for country B is 25% above the mean, with a threshold of 20% around the mean. Both country A and B are then counted equally as 'above average'.

This transformation is used to calculate the summary innovation index (EC - DG ENTR, 2001) in the context of the European Innovation Scoreboard. This index is calculated by the Directorate General Enterprise of the European Commission. Here the component indicators are normalised according to distance from the overall European mean. The summary innovation index is equal to the number of indicators that are at least 20% above the European overall mean, minus the number that are more than 20% below. The index is adjusted for differences in the number of available indicators for each country. The index can vary between +10 (all indicators are above average) to -10 (all indicators are below average).

For time – dependent studies, in order to assess country performance across years, the average across countries $x_{qc=\bar{c}}^{t_0}$ is calculated for a reference year (usually the initial time point t_0). An

indicator that moves from significantly below the mean to significantly above the threshold in the consecutive year will have a positive effect on the composite.

5.2.7 Methods for Cyclical Indicators

Most institutes conducting business tendency surveys select a set of survey series and combine them into cyclical composite indicators. This is done in order to reduce the risk of false signals, and to better forecast cycles in economic activities (Nilsson, 2000).

When indicators are in the form of time series the transformation can be made by subtracting the mean over time $E_t\left(x_{qc}^t\right)$ and then by dividing by the mean of the absolute values of the difference from the mean.

$$I_{qc}^t = \frac{x_{qc}^t - E_t\left(x_{qc}^t\right)}{E_t\left(\left|x_{qc}^t - E_t\left(x_{qc}^t\right)\right|\right)}$$

The normalized series are then converted into index form by adding 100.
This approach is used in the composite leading indicators calculated by the OECD where it is necessary to minimize the influence of series with marked cyclical amplitude to dominate the composite indicator.

The method of normalisation used in the economic sentiment indicators calculated by the Directorate General Economic and Financial Affairs of the European Commission (EU-2004a) consists in transforming the indicator series so that the average month-to-month changes are equal for all the indicators. This treatment is also called **balance of opinions** because, for each indicator, managers of firms from different sectors and sizes are asked to express their opinion upon the firms which have improved and the firms which have reported deterioration with respect to the previous survey. The transformed indicator varies, by construction, between –100 (if all firms have reported deterioration) and +100 (if all firms have noted an improvement).
This method gives implicitly less weight to the more irregular series in the cyclical movement of the composite indicator, unless some prior ad-hoc smoothing is performed.

5.2.8 Percentage of annual differences over consecutive years

Each indicator is transformed using the formula:

$$I_{qc}^t = \frac{x_{qc}^t - x_{qc}^{t-1}}{x_{qc}^t} * 100$$

The transformed indicator is dimension-less. It does represent the percentage growth with respect to the previous year instead of the absolute level. The transformation can be used only when the indicators are available for a number of years.
The method has been applied by the Directorate General Internal Market of the European Commission for the development of the Internal Market Index (Internal Market Scoreboard issue 9, 2001).

Examples of the above transformations are shown in Table 5.6 using the TAI data. The data are sensitive to the choice of the transformation and this might cause problems in terms of loss of the

interval level of the information, sensitivity to outliers, arbitrary choice of categorical scores and sensitivity to weighting.

Table 5.5 *Summary of normalisation methods. Notes:* x_{qc}^{t} *is the value of indicator q for country c at time t.* $\bar{c}$ *is the reference country. The operator sgn gives the sign of the argument (i.e. +1 if the argument is positive, -1 if the argument is negative).* N_e *is the total number of experts surveyed.*

Method	Equation
1. Ranking	$I_{qc}^{t} = Rank(x_{qc}^{t})$
2. Standardization (or z-scores)	$I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc=\bar{c}}^{t}}{\sigma_{qc=\bar{c}}^{t}}$
3. Re-scaling	$I_{qc}^{t} = \frac{x_{qc}^{t} - min_c(x_q^{t_0})}{max_c(x_q^{t_0}) - min_c(x_q^{t_0})}.$
4. Distance to reference country	$I_{qc}^{t} = \frac{x_{qc}^{t}}{x_{qc=\bar{c}}^{t_0}}$ or $I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc=\bar{c}}^{t_0}}{x_{qc=\bar{c}}^{t_0}}$
5. Logarithmic transformation	$I_{qc}^{t} = ln(x_{qc}^{t})$
6. Categorical scales	***if*** x_{qc}^{t} ***in the upper 5 - th percentile then*** $y_{qc}^{t} = 100$ ***if*** x_{qc}^{t} ***in the upper 15 - th percentile then*** $y_{qc}^{t} = 80$ ***if*** x_{qc}^{t} ***in the upper 35 - th percentile then*** $y_{qc}^{t} = 60$ ...
7. Indicators above or below the mean	***if*** $x_{qc}^{t} / x_{qc=\bar{c}}^{t_0} > (1+p)$ ***then*** $I_{qc}^{t} = 1$ ***if*** $x_{qc}^{t} / x_{qc=\bar{c}}^{t_0} < (1-p)$ ***then*** $I_{qc}^{t} = -1$ ***if*** $(1-p) < x_{qc}^{t} / x_{qc=\bar{c}}^{t_0} < (1+p)$ ***then*** $I_{qc}^{t} = 0$
8. Cyclical indicators (OECD)	$I_{qc}^{t} = \frac{x_{qc}^{t} - E_t(x_{qc}^{t})}{E_t(\lvert x_{qc}^{t} - E_t(x_{qc}^{t}) \rvert)}$
9. Balance of opinions (EC)	$I_{qc}^{t} = \frac{100}{N_e} \sum_{e}^{N_e} sgn_e(x_{qc}^{t} - x_{qc}^{t-1})$
10. Percentage of annual differences over consecutive years	$I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc}^{t-1}}{x_{qc}^{t}}$

Table 5.6 *Different normalisation techniques using the TAI data.*

	Mean years of school	Rank	z-score	re-scaling	distance to reference country					Log 10	above/	Percentile	Categorical
	age 15 and above	high value = top in the list			ratio			difference			below		scale
Country					c=mean	c=best	c= worst	c=mean	c=worst		the mean		
											p=20%		
Finland	10	15	0.26	0.59	1.04	0.83	1.41	0.04	0.41	1.00	0	65.2	60
United States	12	23	1.52	1.00	1.25	1.00	1.69	0.25	0.69	1.08	1	100	100
Sweden	11.4	19	1.14	0.88	1.19	0.95	1.61	0.19	0.61	1.06	0	82.6	60
Japan	9.5	12	-0.06	0.49	0.99	0.79	1.34	-0.01	0.34	0.98	0	52.2	50
Korea, Rep. of	10.8	17	0.76	0.76	1.13	0.90	1.52	0.13	0.52	1.03	0	73.9	60
Netherlands	9.4	9	-0.12	0.47	0.98	0.78	1.32	-0.02	0.32	0.97	0	39.1	50
UK	9.4	9	-0.12	0.47	0.98	0.78	1.32	-0.02	0.32	0.97	0	39.1	50
Canada	11.6	20	1.27	0.92	1.21	0.97	1.63	0.21	0.63	1.06	1	87.0	80
Australia	10.9	18	0.83	0.78	1.14	0.91	1.54	0.14	0.54	1.04	0	78.3	60
Singapore	7.1	1	-1.58	0.00	0.74	0.59	1.00	-0.26	0.00	0.85	-1	4.3	0
Germany	10.2	16	0.38	0.63	1.06	0.85	1.44	0.06	0.44	1.01	0	69.6	60
Norway	11.9	22	1.46	0.98	1.24	0.99	1.68	0.24	0.68	1.08	1	95.7	100
Ireland	9.4	9	-0.12	0.47	0.98	0.78	1.32	-0.02	0.32	0.97	0	39.1	50
Belgium	9.3	8	-0.19	0.45	0.97	0.78	1.31	-0.03	0.31	0.97	0	34.8	40
New Zealand	11.7	21	1.33	0.94	1.22	0.98	1.65	0.22	0.65	1.07	1	91.3	80
Austria	8.4	6	-0.76	0.27	0.88	0.70	1.18	-0.12	0.18	0.92	0	26.1	40
France	7.9	5	-1.08	0.16	0.82	0.66	1.11	-0.18	0.11	0.90	0	21.7	40
Israel	9.6	14	0.00	0.51	1.00	0.80	1.35	0.00	0.35	0.98	0	60.9	50
Spain	7.3	4	-1.46	0.04	0.76	0.61	1.03	-0.24	0.03	0.86	-1	17.4	40
Italy	7.2	3	-1.52	0.02	0.75	0.60	1.01	-0.25	0.01	0.86	-1	13.0	20
Czech Republic	9.5	12	-0.06	0.49	0.99	0.79	1.34	-0.01	0.34	0.98	0	52.2	50
Hungary	9.1	7	-0.31	0.41	0.95	0.76	1.28	-0.05	0.28	0.96	0	30.4	40
Slovenia	7.1	1	-1.58	0.00	0.74	0.59	1.00	-0.26	0.00	0.85	-1	4.3	0

Sometimes, there is no need to carry out a normalisation of the indicators. For example, if the indicators are already expressed with the same standard. See, for example, the case of the e-business readiness composite indicator (Nardo et al., 2004). Here, all the sub-indicators are expressed in terms of percentage of enterprises possessing a given infrastructure or using a given ICT tool. In such case, the normalization would rather obfuscate the issue, as one would loose the inherent information contained in the percentages.

6. Weighting and Aggregation

Central to the construction of a composite index is the need to combine in a meaningful way different dimensions measured on different scales. This implies a decision on which weighting model will be used and which procedure will be applied to aggregate the information.

Different weights may be assigned to component series in order to reflect their economic significance (collection costs, coverage, reliability and economic reason), statistical adequacy, cyclical conformity, speed of available data, etc. In this section a number of techniques are presented ranging from weighting schemes based on statistical models (such as factor analysis, data envelopment analysis, unobserved components models), to participatory methods (e.g. budget allocation or analytic hierarchy processes). Weights usually have an important impact on the value of the composite and on the resulting ranking especially whenever higher weight is assigned to sub-indicators on which some countries excel or fail. This is why weighting models need to be made explicit and transparent. Moreover, the reader should bear in mind that, no matter which method is used, weights are essentially value judgments and have the property to make explicit the objectives underlying the construction of a composite (Rowena et al., 2004).
Weighting is strongly related to how the information conveyed by the different dimensions is aggregated into a composite index. Different aggregation rules are possible. Sub-indicators could be summed up, multiplied or aggregated using non linear techniques. Each technique implies different assumptions and has specific consequences. This section revises the main methods for aggregating sub-indicators into a composite index. However, since several variations on each method exist, this review does not pretend to be comprehensive but rather to supply the reader with a critical assessment of the most common methodologies.

The layout of the section is the following. The first part is devoted to the issue of weighting sub-indicators while the second part deals with the aggregation of the (weighted) sub-indicators into a composite index. A succinct "when to use what" checklist concludes.

6.1 Weighting

No agreed methodology exists to weight individual indicators. An analyst might be willing to reward with higher weight the components that are deemed more influential, regardless of any other consideration. Another might pay great attention to the existence of correlations among factors or use weights derived from principal components analysis to overcome the double counting problems when two or more indicators partially measure the same behaviour. Indicators could also be weighted based on the opinion of experts, who know policy priorities and theoretical backgrounds, to reflect the multiplicity of stakeholders' viewpoints.

Weights heavily influence the outcome of a composite indicator and countries ranking in a benchmarking exercise. Therefore, weights should ideally be selected according to an underlying and agreed or at least clearly stated theoretical framework. Weighting imply a "subjective" evaluation, which is particularly delicate in case of complex, interrelated and multidimensional phenomena.

Indicators, and a fortiori composite indicators, are models, similar in their nature and in the way they are encoded, to mathematical or computational models, such as those created to describe the spread of diseases, the movement of tides, the production of a chemical plant, the cycles of the economy. Exactly as in these examples, no formal encoding procedure exists, relating the process

being modelled to its representation, rather than the modeller craftsmanship, and a justification of the practice lays in its fitness to the intended purpose (Rosen, 1991).

Whatever method is used to derive weights, no consensus is likely to exist. This should not preclude use of a composite, but highlights the dangers of presenting any composite as "objective". At best, it indicates a set of priorities that has been informed by popular or expert judgments (including the analyst). Assumptions ad implication of the used weighting system should be always made clear and tested for robustness. Soundness and transparency should guide the entire exercise.

In many composite indicators all variables are given the same weight when there are no statistical or empirical grounds for choosing a different scheme. **Equal weighting** (EW) could imply the recognition of an equal status for all sub-indicators (e.g. when policy assessments are involved). Alternatively, it could be the result of insufficient knowledge of causal relationships, or ignorance about the correct model to apply (like in the case of Environmental Sustainability Index - World economic forum, 2002), or even stem from the lack of consensus on alternative solutions (as happened with the Summary Innovation Index - European Commission, 2001a). In any case, EW does not mean no weighting, because EW anyway implies an implicit judgment on the weights being equal. The effect of EW also depends on how component indicators are divided into categories or groups: weighting equally categories regrouping a different number of sub-indicator could disguise different weights applied to each single sub-indicator.

Weights may also reflect the statistical quality of the data, thus higher weight could be assigned to statistically reliable data (data with low percentages of missing values, large coverage, sound values). In this case the concern is to reward only easy to measure and readily available base-indicators, punishing the information that is more problematic to identify and measure.

Weights based on statistical models

When using equal weighting it may happen that - by combining variables with high degree of correlation – one may introduce an element of double counting into the index: if two collinear indicators are included in the composite index with a weight of w_1 and w_2, than the unique dimension that the two indicators measure would have weight (w_1+w_2) in the composite. The response has often been testing indicators for statistical correlation - for example with the Pearson correlation coefficient (Manly, 1994) - and choosing only indicators exhibiting a low degree of correlation or adjusting weights correspondingly, e.g. giving less weight to correlated indicators. Furthermore, minimizing the number of variables in the index may be desirable on other grounds such as transparency and parsimony.

Notice that there will almost always be some positive correlation between different measures of the same aggregate. Thus, a rule of thumb should be introduced to define a threshold beyond which the correlation is a symptom of double counting. On the other hand relating correlation analysis to weighting could be dangerous when motivated by apparent redundancy. For example, in the CI of e-business readiness the indicator I_1 "Percentage of firms using Internet" and indicator I_2 "The percentage of enterprises that have a web site" display a correlation of 0.88 in 2003: are we allowed to give less weight to the pair (I_1, I_2) given the high correlation or shall we consider the two indicators as measuring different aspects of Innovation and Communication Technologies Adoption and give them equal weight in constructing the composite indicator? If weights should ideally reflect the contribution of each indicator to the composite, double counting

should not only be determined by statistical analysis but also by the analysis of the indicator itself vis à vis the rest of indicators and the phenomenon they all aim to picture.

6.1.1 Principal component analysis and factor analysis

Principal component analysis (PCA) and more specifically factor analysis (FA) (Section 3) group together sub-indicators that are collinear to form a composite indicator capable of capturing as much of common information of those sub-indicators as possible. The information must be comparable for this approach to be used: sub-indicators must have the same unit of measurement. Each factor (usually estimated using principal components analysis) reveals the set of indicators having the highest association with it. The idea under PCA/FA is to account for the highest possible variation in the indicators set using the smallest possible number of factors. Therefore, the composite no longer depends upon the dimensionality of the dataset but it is rather based on the "statistical" dimensions of the data. According to PCA/FA, weighting only intervenes to correct for the overlapping information of two or more correlated indicators, and it is not a measure of importance of the associated indicator.

If no correlation between indicators is found, then weights can not be obtained estimated with this method. This is the case for the new economic sentiment indicator, where factor and principal components analysis excluded the weighing of individual questions within a sub-component of the composite index (see the supplement B of the Business and Consumer Surveys Result N. 8/9 August/September 2001[13]). PCA/FA was excluded in the construction of an indicator of environmental sustainability when it was found that this procedure assigned negative weights to some sub-indicators (World Economic Forum, 2002).

Methodology

The first step in FA is to check the correlation structure of the data: if the correlation between the indicators is low then it is unlikely that they share common factors.

The second step is the identification of a certain number of latent factors, small than the number of sub-indicators, representing the data. Summarizing briefly what has been explained in Section 3, each factor depends on a set of coefficients (loadings), each coefficient measuring the correlation between the individual indicator and the latent factor. Principal component analysis is usually used to extract factors (Manly, 1994[14]). For a factor analysis only a subset of principal components are retained (let's say m), the ones that account for the largest amount of the variance.

The standard practice is to choose factors that: (i) have associated eigenvalues larger than one; (ii) individually contribute to the explanation of overall variance by more than 10%; (iii) cumulatively contribute to the explanation of the overall variance by more than 60%. With the TAI reduced dataset (the one with 23 countries) the factors with eigenvalues close to the unity are the first four, as summarized in Table 6.1. Individually they explain more than 10% of the total variance and overall they count for about the 87% of variance.

[13] http://europa.eu.int/comm/economy_finance/publications/european_economy/2001/b2001_0809_en.pdf)

[14] Other methods are available, e.g. the Maximum Likelihood or the principal Factor centroids. Notice that these methods usually supply very different weights especially when the sample size of FA is small.

Table 6.1. *Eigenvalues of the Technology Achievement Index. Dataset with 23 countries.*

	Eigenvalues	% Total variance	Cumulative (%)
1	**3.3**	**41.9**	41.9
2	**1.7**	**21.8**	63.7
3	**1.0**	**12.3**	76.0
4	**0.9**	**11.1**	**87.2**
5	0.5	6.0	93.2
6	0.3	3.7	96.9
7	0.2	2.2	99.1
8	0.1	0.9	100.0

The third step involves the rotation of factors. The rotation (usually the *varimax rotation*) is used to minimize the number of sub-indicators that have a high loading on the same factor. The idea in transforming the factorial axes is to obtain a "simpler structure" of the factors (ideally a structure in which each indicator is loaded exclusively on one of the retained factors). Rotation is a standard step in factor analysis, it changes the factor loadings and hence the interpretation of the factors leaving unchanged the analytical solutions obtained *ex-ante* and *ex-post* the rotation.

Table 6.2. *Factor loadings of Technology Achievement Index. Varimax normalised, extraction: Principal Components.*

	Factor 1	Factor 2	Factor 3	Factor 4	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*
Patents	0.07	0.97	0.06	0.06	*0.00*	***0.68***	*0.00*	*0.00*
Royalties	0.13	0.07	-0.07	0.93	*0.01*	*0.00*	*0.00*	***0.49***
Internet	0.79	-0.21	0.21	0.42	***0.24***	*0.03*	*0.04*	*0.10*
Tech exports	-0.64	0.56	-0.04	0.36	*0.16*	***0.23***	*0.00*	*0.07*
Telephones	0.37	0.17	0.38	0.68	*0.05*	*0.02*	*0.12*	***0.26***
Electricity	0.82	-0.04	0.25	0.35	***0.25***	*0.00*	*0.05*	*0.07*
Schooling	0.88	0.23	-0.09	0.09	***0.29***	*0.04*	*0.01*	*0.00*
University	0.08	0.04	0.96	0.04	*0.00*	*0.00*	***0.77***	*0.00*
Expl.Var	2.64	1.39	1.19	1.76				
Expl./Tot	*0.36*	*0.26*	*0.24*	*0.42*				

The last step deals with the construction of the weights from the matrix of factor loadings after rotation, given that the square of factor loadings represent the proportion of the total unit variance of the indicator which is explained by the factor. The approach used by Nicoletti G., Scarpetta S., Boylaud O. (2000) is that of grouping the sub-indicators with the highest factors loadings in *intermediate* composite indicators. With the TAI dataset the *intermediate* composites are 4 (Table 6.2). The first includes Internet (with a weight of 0.24), Electricity (weight 0.25) and Schooling (weight 0.29).[15] Likewise the second *intermediate* is formed by Patents and Technology Exports (worth 0.68 and 0.23 respectively), the third only by University (0.77) and the fourth by Royalties and Telephones (weighted with 0.49 and 0.26).

[15] Weights are normalized squared factor loading, e.g. 0.24 = (0.79^2)/2.64 which is the portion of the variance of the first factor explained by the variable Internet.

Then the four *intermediate* composites are aggregated by weighting each composite using the proportion of the explained variance in the dataset: 0.36 for the first (0.36 = 2.64/(2.64+1.39+1.19+1.76)), 0.26 for the second, 0.24 for the third and 0.42 for the fourth.[16]
Notice that different methods for the extraction of principal components imply different weights, hence different scores for the composite (and possibly different country ranking). For example if Maximum Likelihood (ML) were to be used instead of Principal Component (PC) the weights obtained would be:

	M L	PCA
Patents	0.19	0.17
Royalties	0.20	0.20
Internet	0.07	0.08
Tech exports	0.07	0.06
Telephones	0.15	0.11
Electricity	0.11	0.09
Schooling	0.19	0.10
University	0.02	0.18

PCA/FA for weighting indicators

Advantages

- It does not imply any manipulation of weights through *ad hoc* restrictions.
- It solves the double counting problem

Disadvantages

- It can only been used with correlated sub-indicators.
- Sensitive to modifications of basic data: data revisions and updates (e.g. new observations and new countries) may change the set of weights (i.e. the estimated loadings) used in the composite.
- Sensitive to the presence of outliers, that may introduce spurious variability in the data
- Sensitive to small-sample problems and data shortage that may make the statistical identification or the economic interpretation difficult (in general a relation between data and unknown parameters of 3:1 is required for a stable solution).
- Minimize the contribution of indicators, which do not move with other indicators.
- Sensitive to the factor extraction and to the rotation methods.

Examples of use

Indicators of product market regulation (Nicoletti et al., OECD, 2000)
Internal Market Index (EC-DG MARKT, 2001b)
Business Climate Indicator (EC-DG ECFIN, 2000)
General Indicator of S&T (NISTEP, 1995)
Success of software process Improvement (Emam et al. 1998)

[16] To preserve comparability final weights could be rescaled to sum up to one.

6.1.2 Data envelopment analysis and Benefit of the doubt

Data envelopment analysis (DEA) employs linear programming tools (popular in Operative Research) to retrieve an *efficiency frontier* and uses this as benchmark to measure the performance of a given set of countries.[17] The set of weighs stems from this comparison. Two main issues are involved in this methodology: the construction of a benchmark (the frontier) and the measurement of the distance between countries in a multi-dimensional framework.

The construction of the **benchmark** is done by assuming:

(i) positive weights (the higher the value of one sub-indicator, the better for the corresponding country);

(ii) non discrimination of countries that are best in any single dimension (i.e. sub-indicator) thus ranking them equally; and

(iii) a linear combination of the best performers is feasible (convexity of the frontier).

The distance of each country with respect to the benchmark is determined by the location of the country and its position relative to the frontier. Both issues are represented in Figure 6.1, for the simple case of 4 countries and only two base indicators.

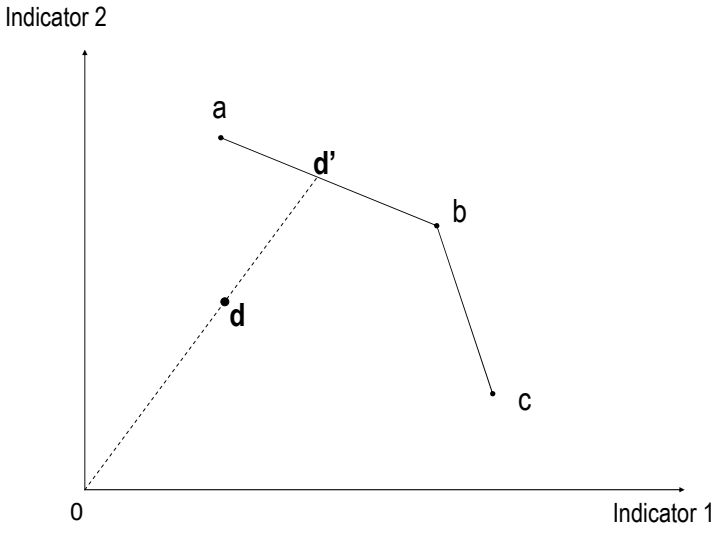


Figure 6.1. *Performance frontier determined with Data Envelopment Analysis. Rearranged from Mahlberg and Obersteiner (2001).*

In Figure 6.1 two indicators are represented in the two axes and four countries (*a, b, c, d*) are ranked according to the score of the indicators. The line connecting countries *a*, *b* and *c* constitutes the performance frontier and is the benchmark for country *d* which lies beyond the frontier. The countries supporting the frontier are classified as the best performing, while country *d* is the worse performing. The performance indicator is the ratio of the distance between the

[17] DEA has also been used in production theory, for a review see Charnes et al, 1995.

origin and the actual observed point and that of the projected point in the frontier: $\overline{0d} / \overline{0d'}$. The best performing countries will have a performance score of 1, while for the least performing it will be lower than one. This ratio corresponds to the expression $(w_{1d} I_{1d} + w_{2d} I_{2d}) / (w_{1d} I^*_{1d} + w_{2d} I^*_{2d})$, where I^*_{id} is the frontier value of indicator *i*, *i=1,2*, while I_{id} is its actual value (see expression 6.1 for more than 2 indicators). The set of weights of each country will therefore depend on its position with respect to the frontier. The benchmark will correspond to the ideal point exhibiting a similar mix of indicators (*d'* in the example).

The benchmark could also be determined by a hypothetical decision maker (Korhonen et al. 2001, for an indicator of performance of academic research) who is asked to locate the target in the efficiency frontier having the most preferred combination of sub-indicators. In this case the DEA approach could merge with the budget allocation method (see below) since experts are asked to assign weights (i.e. priorities) to sub-indicators.

Benefit of the doubt approach

This methodology, originally proposed for evaluating macroeconomic performance (Melyn and Moesen, 1991) and only recently adapted to the index theory[18], is an application of the DEA. The composite indicator is defined as the ratio of a country's actual performance over its benchmark performance:

$$CI_c = \frac{\sum_{c=1}^{M} I_{qc} w_{qc}}{\sum_{c=1}^{M} I^*_{qc} w_{qc}} \qquad (6.1)$$

Where I_{qc} is the normalized (with the max-min method) score of q^{th} sub-indicator ($q=1,\ldots,Q$) for country *c* ($c=1,\ldots M$) and w_{qc} the corresponding weight.

Cherchye et al. (2004) who first implemented this method suggested obtaining the benchmark as solution of a maximization problem (although external benchmarks are also possible):

$$I^* = I^*(w) = \underset{I_k, k \in \{1,\ldots,M\}}{arg\,max} \left(\sum_{q=1}^{Q} I_{qk} w_q \right) \qquad (6.2)$$

I^* is the score of the hypothetical country that maximizes the overall performance (defined as the weighted average), given the (unknown) set of weights *w*. Notice that (i) weights are country specific: different sets of weights may lead to choose different countries as far as there is no country having the highest score in all sub-indicators; (ii) the benchmark would in general be country-dependent, so no unique benchmark would exist (unless, as before, a country is better-off in all sub-indicators), (iii) sub-indicators must be comparable, i.e. have the same unit of measurement.

The second step is the specification of the set of weights for each country. The optimal set of weighs (if it exists) guarantees the best position for the associated country *vis á vis* all other

[18] We present the method as it has been used in Cherchye et al. 2003, and Cherchye and Kuosmanen 2002.

countries in the sample: with any other weights profile the relative position of that country would have been worse. Optimal weights are obtained by solving the following problem:

$$CI_c^* = \underset{w_{qc},q=1,\ldots,Q}{arg\,max} \frac{\sum_{q=1}^{Q} I_{qc} w_{qc}}{\underset{I_k,k\in\{1,\ldots,M\}}{max}\left(\sum_{q=1}^{Q} I_{qk} w_{qc}\right)} \qquad \text{for } c=1,..,M \tag{6.3}$$

subject to non negativity constraints on weights. [19] The resulting composite index will range between zero (lowest possible performance) and 1 (the benchmark). Operationally, expression (6.3) can be reduced to the linear programming problem (6.4) by multiplying all weights by a common factor (that does not alter the index value) and solved using optimizations algorithms

$$\begin{aligned} & CI_c^* = \underset{wqc}{arg\,max} \sum_{q=1}^{Q} I_{qc} w_{qc} \\ & s.t. \\ & \sum_{q=1}^{Q} I_{qk} w_{qk} \le 1 \\ & w_{qk} \ge 0 \\ & \forall k = 1,\ldots,M; \forall q = 1,\ldots,Q \end{aligned} \tag{6.4}$$

The results of the benefit of the doubt (BOD) approach applied to the TAI example can be seen in Table 6.3. Weights are in the first eight columns wile the last column contains the composite indicator values. The example deserves a couple of remarks. The first is that Finland, USA and Sweden have a composite index value equal to one, i.e. they all score first in the ranking. This however hides a problem of multiplicity of equilibria: in Figure 6.1 any point between country *a* (say Finland) and country *b* (say USA) can be an optimal solution for these countries. Thus weights are not uniquely determined (although if the CI is unique). The weights values for these three countries given in Table 6.3 are thus only three among many (infinite) possible weighting schemes. Notice also that the multiplicity of solutions is likely to depend upon the set of constraints imposed to the weights of the maximization problem in (6.4): the wider is the range of variation of weights and the lower is the possibility of obtaining a unique solution[20].

Second, the set of weights for each country, as calculated by the above algorithm, that does not sum up to one, making the comparison with other methods (like FA or EW) impossible.[21] A rescaling to assure comparability would be innocuous only in the case of a unique solution for all countries. If multiplicity arise, instead, the scaling to a unit interval may be arbitrary.

[19] Additional constraints could be imposed. Country-specific restrictions to reflect prior information can also be added.
[20] In our example we imposed the requirement for each sub-indicator to weight at least 10% and no more than 15% of the total.
[21] However, precisely since the BOD weights have the meaning of relative weights, the weights as originally produced by the algorithm can always be normalized afterwards so as to sum up to one, leaving its intrinsic meaning unaffected and at the same time facilitating comparison with other methods' results.

Table 6.3. *BOD approach applied to the TAI dataset (23 countries). Columns 1 to 8 contain weights, column 9 displays the country's composite indicator.*

	Patents	Royalties	Internet	Tech. Exports	Telephones	Electricity	Schooling	University	CI
Finland	0.15	0.17	0.17	0.16	0.19	0.17	0.17	0.19	**1**
United States	0.20	0.20	0.17	0.21	0.15	0.15	0.21	0.14	**1**
Sweden	0.18	0.21	0.15	0.19	0.19	0.16	0.20	0.14	**1**
Japan	0.22	0.15	0.15	0.22	0.22	0.16	0.21	0.15	**0.87**
Korea	0.22	0.14	0.14	0.22	0.14	0.14	0.22	0.22	**0.80**
Netherlands	0.22	0.22	0.14	0.22	0.22	0.14	0.14	0.14	**0.75**
United Kingdom	0.14	0.21	0.14	0.21	0.21	0.14	0.20	0.15	**0.71**
Canada	0.14	0.14	0.14	0.21	0.21	0.21	0.21	0.14	**0.73**
Australia	0.13	0.13	0.20	0.13	0.13	0.20	0.20	0.20	**0.66**
Singapore	0.14	0.14	0.14	0.20	0.20	0.20	0.14	0.20	**0.62**
Germany	0.22	0.15	0.15	0.22	0.21	0.15	0.22	0.15	**0.62**
Norway	0.14	0.14	0.20	0.14	0.20	0.20	0.20	0.14	**0.86**
Ireland	0.14	0.21	0.14	0.21	0.21	0.14	0.20	0.15	**0.60**
Belgium	0.14	0.16	0.14	0.21	0.19	0.21	0.21	0.14	**0.54**
New Zealand	0.21	0.14	0.21	0.14	0.14	0.21	0.21	0.14	**0.58**
Austria	0.22	0.14	0.14	0.22	0.22	0.22	0.14	0.14	**0.52**
France	0.22	0.14	0.14	0.22	0.22	0.22	0.14	0.14	**0.51**
Israel	0.21	0.15	0.15	0.22	0.22	0.15	0.22	0.15	**0.49**
Spain	0.21	0.14	0.14	0.21	0.21	0.14	0.14	0.21	**0.34**
Italy	0.22	0.14	0.14	0.22	0.22	0.22	0.14	0.14	**0.38**
Czech Rep.	0.22	0.15	0.15	0.22	0.15	0.22	0.22	0.15	**0.31**
Hungary	0.22	0.14	0.21	0.22	0.14	0.14	0.22	0.15	**0.27**
Slovenia	0.22	0.14	0.14	0.22	0.22	0.22	0.14	0.14	**0.28**

Benefit of the Doubt

Advantages	**Disadvantages**
▪ The indicator will be sensible to national policy priorities, in that the weights are endogenously determined by the observed performances (this is a useful second best approach whenever the first best – full information about true policy priorities- can not be attained).	▪ Weights are country specific, thus cross-country comparisons is not possible. ▪ Without imposing constraints on weights (except the non-negativity) the most likely solution is to have all countries with a composite equal to 1. When constraints on weights are imposed it may be the case that, for

▪ The benchmark is not based upon theoretical bounds but it a linear combination of observed best performances. ▪ It is useful in policy arena, since policy makers could not complain about unfair weighting: any other weighting scheme would have generated lower composite scores. ▪ Such an index could be "incentive generating" rather than "punishing" the countries lagging behind. ▪ Weights, by revealing information about the policy priorities, may help to define trade-offs, overcoming the difficulties of linear aggregations.	some country, no solution of the maximization problem exist, likewise it may happen that there exist a multiplicity of solutions making the optimal set of weights undetermined (this is likely to happen when the *CI*=1). ▪ Different normalizations of the scores are likely to give different weighting schemes. ▪ The index is likely to reward the status-quo, since for each country the maximization problem gives higher weights to higher scores. ▪ Endogenous weighting has the risk of substituting open experts' opinions with the analyst's manipulation of weights (through the constraints). Transparency of the procedure would be lost. ▪ The value of the scoreboard depends on the benchmark performance. If this changes the composite will change as well as the set of weights (and the country ranking). ▪ The best performer (the one with a composite equal to one) will not see its progress reflected in the composite (that will remain stacked to 1). This can be solved by imposing an external benchmark.

Examples of use

Human Development Index (Mahlberg and Obersteiner, 2001)
Sustainable development (Cherchye and Kuosmanen, 2002)
Social Inclusion (Cherchye, Mosen, Van Puyenbroeck, 2004)
Macro-economic performance evaluation (Melynand and Moesen, 1991, and Cherchye 2001)
Unemployment (Storrie and Bjurek, 1999, and 2000)

6.1.3 Regression approach

Linear regression models can tell us something about the 'linkages' between a large number of indicators I_{1c}, I_{2c}, ..., I_{Qc} and a single output measure $\hat{Y}_c$ representing the objective to be attained. A (usually linear) multiple regression model is then estimated to retrieve the relative weights of sub-indicators:

$$\hat{Y}_c = \hat{\alpha} + \hat{\beta}_1 I_{1c} + ... + \hat{\beta}_Q I_{Qc} \qquad (6.5)$$

where $\hat{Y}_c$, $c=1,...,M,$ is a measure (not necessarily an indicator) of the phenomenon that sub-indicators aim to picture, $\hat{\alpha}$ is the estimated constant and $\hat{\beta}_1$ to $\hat{\beta}_Q$ are the regression coefficients (weights) of the associated sub-indicators I_1, I_2,..., I_Q.

This approach, although suitable for a large number of variables of different types, implies the assumption of linear behaviour and requires the independence of explanatory variables. If these variables are correlated, in fact, estimators will have high variance meaning that parameters estimates will not be precise and hypothesis testing not powerful. In the extreme case of perfect collinearity among regressors the model will not even be identified. It is further argued that if the concepts to be measured could be represented by a single measure $\hat{Y}_c$, then there would be no need for developing a composite indicator (*Muldur 2001*). Yet this approach could still be useful to verify and adjust weights, or when interpreting sub-indicators as possible policy actions. The regression model, thereafter, could quantify the relative effect of each policy action on the target, i.e. a suitable output performance indicator identified on a case-by-case basis.

Regression Approach

Advantages	**Disadvantages**
▪ It can be used even if component indicators are not correlated. ▪ It does not imply any manipulation of weights through ad hoc restrictions. ▪ It is useful to update or validate the applied set of weights.	▪ It provides poor results in case of highly correlated component indicators (multi-collinearity problems). Remedies can be found associating PCA with regression analysis. ▪ It requires a large amount of data to produce estimates with known statistical properties.

Examples of use

Composite Economic Sentiment Indicator (ESIN) http://europa.eu.int/comm/economy_finance
National Innovation Capacity index (Porter and Stern, 1999)

6.1.4 Unobserved components models

Weights with the unobserved components models (UCM) are obtained by estimating with the maximum likelihood method a function of the base indicators. The idea is that sub-indicators depend on an unobserved variable plus an error term, e.g. the "percentage of firms using internet in country *j*" depends upon the (unknown) propensity to adopt new information and communication technologies plus an error term accounting, for example, for the error in the sampling firms. Therefore, by estimating the unknown component it will possible to shed some light on the relationship between the composite and its components. The weight obtained will be set so as to minimize the error in the composite. This method resembles the previous one (even if with a different interpretation). The main difference resides in the dependent variable that with UCM is unknown.

Methodology

To use this method one needs a set of sub-indicators, all measuring an unknown phenomenon (e.g. the 8 sub-indicators of TAI measuring the capacity of a country to participate in the "network age"). Let *ph(c)* the unknown phenomenon to be measured. The observed data consist on a cluster of *q=1,...,Q(c)* indicators, each measuring an aspect of *ph(c)*. Let *c=1,...M(q)* the countries covered by indicator *q*. The observed score of country *c* on indicator *q*, *I(c,q)*, can be written as a linear function of the unobserved phenomenon and of an error term, $\varepsilon(c,q)$:

$$I(c,q) = \alpha(q) + \beta(q)[\,ph(c) + \varepsilon(c,q)\,] \qquad (6.6)$$

$\alpha(q)$ and $\beta(q)$ are unknown parameters mapping *ph(c)* on *I(c,q)*.
The error term captures two sources of uncertainty in the relationship between the phenomenon and one of its indicators. First the phenomenon could be only imperfectly measured or observed in each country (e.g. because of errors of measurement). Second the relationship between *ph(c)* and *I(c,q)* is imperfect (e.g. *I(c,q)* may only be a noisy indicator of the phenomenon if there are differences among countries in what the indicator is considered to be). The error term $\varepsilon(c,q)$ is assumed to have zero mean, $E(\varepsilon(c,q))=0$, and the same variance across countries within a given indicator (but a different variance across indicators), $E(\varepsilon(c,q)^2)=\sigma_q^2$; it also holds $E(\varepsilon(c,q)\varepsilon(i,h))=0$ for $c \neq i$ or $q \neq h$.

The assumption that errors are independent across indicators is based on the idea that sub-indicators should ideally give independent information about one particular aspect of the phenomenon. Dropping this assumption is rather complicate since it would imply separating the correlation due to the collinearity of indicators from the correlation of error terms in order to obtain sound estimates.

Furthermore, to facilitate calculations it is usually assumed that *ph(c)* is a random variable with mean zero and unit variance and the indicators are rescaled to take values between zero and one. The assumption that both *ph(c)* and $\varepsilon(c,q)$ are jointly normally distributed simplifies the estimation of the level of *ph(c)* in country *c*, which is done by using the mean of the conditional distribution of the unobserved component (once the observed scores are appropriately rescaled):

$$E[ph(c)/I(c,1),...,I(c,Q(c))] = \sum_{q=1}^{Q(c)} w(c,q)\frac{I(c,q)-\alpha(q)}{\beta(q)} \tag{6.7}$$

The weights are equal to:

$$w(c,q) = \frac{\sigma_q^{-2}}{1+\sum_{q=1}^{Q(c)}\sigma_q^{-2}} \tag{6.8}$$

w(c,q) is a decreasing function of the variance of indicator *q* (expressing the idea that the lower is the precision of indicator *q*, the lower will be the weight assigned to that indicator), and an increasing function of the variance of the other indicators. The weight depends on the country considered in the following way: *w(c,q)* depends on the variance of indicator *q* (numerator) and on the sum of the variances of the all the other sub-indicators including *q* (denominator). However, since not all countries have data on all sub-indicators, the denominator of *w(c,q)* could be made up of a country-dependent number of elements. This may produce non comparability of country values for the composite in the same way as BOD does. Obviously whenever the set of indicators is equal for all countries then weights will no longer be country specific and comparability will be assured. The variance of the conditional distribution is given by:

$$var[ph(c)/I(c,1),...,I(c,Q(c))] = [1+\sum_{q=1}^{Q(c)}\sigma_q^{-2}]^{-1} \tag{6.9}$$

and can be seen as a measure of the precision of the composite indicator useful to construct confidence intervals. This variance is decreasing in the number of indicators for each country and increasing in the variance of the disturbance term for each indicator.

The estimation of the model is made easier by the assumption of normality for *ph(c)* and $\varepsilon(c,q)$. The likelihood function of the observed data is maximized with respect to the unknown parameters, $\alpha(q)s$, $\beta(q)s$, and $\sigma_q^2 s$, and their estimated values substituted in equation (6.7) to obtain the composite indicator and the weights.

Unobserved components model

Advantages	**Disadvantages**
▪ Weights do not depend on ad hoc restrictions.	▪ Reliability and robustness of results depend on the availability of enough data. ▪ With highly correlated sub-indicators there could be identification problems. Thus the method is likely to work well with independent sub-indicators. ▪ The method rewards the absence of outliers, given that weights are a decreasing function of the variance of sub-indicators. ▪ If each country has a different number of sub-indicators; weights are hence country specific.

Examples of use

Governance indicators (see Kaufmann, Kraay and Zoid-lobatón, 1999 and 2003)

6.1.5 Budget allocation

Budget allocation (BAL) is a participatory method in which experts are given a "budget" of *N* points, to be distributed over a number of sub-indicators, "paying" more for those indicators whose importance they want to stress (Moldan and Billharz, 1997). The budget allocation method implies in four different phases:

- Selection of experts for the valuation;
- Allocation of budget to the sub-indicators;
- Calculation of the weights;
- Iteration of the budget allocation until convergence is reached (optional).

It is essential to bring together experts that have a wide spectrum of knowledge, experience and concerns, so as to ensure that a proper weighting system is found for a given application. A case study in which 400 German experts in 1991 were asked to allocate a budget to several environmental indicators related to an air pollution problem showed very consistent results, in spite of the fact that the experts came from opposing social spheres like the industrial sector and the environmental sector (Jesinghaus in: Moldan and Billharz, 1997). Special care should be given in the identification of the population of experts from which to draw a sample, stratified or otherwise.

Budget allocation	
Advantages	**Disadvantages**
▪ Weighting is based on experts' opinion and not on technical manipulations. ▪ Experts' opinions are likely to increase the legitimacy of the composite and create a forum of discussion around which to form a consensus for policy action.	▪ Weighting reliability. Weights could reflect specific local conditions (e.g. in environmental problems), so expert weighting may not be transferable from one area to another. ▪ Allocating a certain budget over a too large number of indicators can give serious cognitive stress to the experts, as it implies circular thinking. The method is likely to produce inconsistencies for a number of indicators higher than 10. ▪ The weighting may not measure the importance of each sub-indicator but rather the urgency or need for political intervention in the dimension of the sub-indicator concerned (e.g. more weight on Ozone emissions if the expert feels that not enough has been made to abate them).
Examples of use	
Employment Outlook (OECD,1999) Composite Indicator on e-Business Readiness (EC-JRC, 2004b). National Health Care System Performance (King's Fund., 2001) Eco-indicator 99 (Pré-Consultants NL, 2000) (weights based on survey from experts). Overall Health System Attainment (WHO, 2000) (weights based on survey from experts)	

6.1.6 Public opinion

Instead of letting experts determine the weights of the indicators in an index, one could ask the general public. Parker (1991, p. 95-98*)* argues that "*public opinion polls have been extensively employed for many years for many purposes, including the setting of weights and they are easy to carry out and inexpensive*". In public opinion polls, issues are selected which are already on the public agenda, and thus enjoy roughly the same attention in the media. From a methodological point of view, opinion polls focus on the notion of "concern", that is people are asked to express "much" or "little concern" about certain problems measured by the base indicators. As with expert assessments, the budget allocation method could also be applied in public opinion polls. However it is more difficult to ask the public to allocate a hundred points to several sub-indicators than to express a degree of concern about the problems that the indicators represent.

Public Opinion	
Advantages	**Disadvantages**
▪ deals with issues on the public agenda. ▪ allows all stakeholders to express their	▪ implies the measurement of "concern" . ▪ the method could produce inconsistencies

preference, and creates a consensus for policy actions.

when dealing with high number of indicators.

Examples of use

Concern about environmental problems Index (Parker, 1991)

6.1.7 Benchmarking with "distance to the target"

One way to avoid the immediate selection of weights is to measure the need for political intervention and the "urgency" of a problem by the distance to target approach. The urgency is high if we are far away from the goal, and low if the goal is almost reached. The weighting itself is realized by dividing the sub-indicator values by the corresponding target values, both expressed in the same units. The dimensionless parameters that are obtained in this way can be summarized by a simple average to produce the composite indicator.

Which target? Policy, sustainability, avoidance of the damage (cost to avoid, notice that this should be the same as the distance to the goal, since the distance to the goal should be proportional to the cost to reach the goal). Alternatively to policy goals, sustainability levels, quantified effects on the environment, or best performance countries can be used as goalposts.

Distance to the Target

Advantages

- The use of policy goals as targets convinces the policy makers for the "soundness" of the weighting method, as long as those policy makers have defined the policy targets themselves.
- This approach is technically feasible when there is a well-defined basis for a certain policy, such as a National Policy Plan or similar reference documents.

Disadvantages

- For international comparisons, such reference policies are often not available, or they deliver contradictory results.
- The benefits of a given policy should be valued independently of the existing policy goals.

Examples of use

- Environmental Policy Performance Indicator (Adriaanse, 1993)
- Human Development Index (UN, 1990, 2000)

6.1.8 Analytic Hierarchy Process

The Analytic Hierarchy Process (AHP) - proposed by Thomas Saaty in the 1970s - is a widely used technique for multi-attribute decision making (Saaty, 1987). It enables the decomposition of a problem into hierarchy and assures that both qualitative and quantitative aspects of a problem are incorporated in the evaluation process, during which opinions are systematically extracted by means of pairwise comparisons. According to Forman et al. (1983): "*AHP is a compensatory decision methodology because alternatives that are efficient with respect to one or more objectives can compensate by their performance with respect to other objectives. AHP allows for the application of data, experience, insight, and intuition in a logical and thorough way within a*

hierarchy as a whole. In particular, AHP as weighting method enables decision-maker to derive weights as opposed to arbitrarily assign them."
The compensatory feature of the method implies that each weights obtained is a trade-offs, i.e. it indicates how much a group of interviewed actors, on average, is willing to forego a given variable in exchange for another variable (for example, in TAI example, how much "patents" can be exchanged for University enrolment). Weights obtained with the AHP are not importance coefficients, i.e. they do not indicate the degree of relevance each alternative has in explaining the phenomenon captured by the composite. This feature has generated many misunderstanding in the literature of composite indicators that often used AHP weights as importance coefficients (see Ülengin et al. 2001).

Methodology

The core of AHP is an ordinal pair-wise comparison of attributes, sub-indicators in this context, in which preference statements are addressed. For a given objective, the comparisons are made per pairs of sub-indicators by firstly posing the question - Which of the two is the more important? - and secondly - By how much? The strength of preference is expressed on a semantic scale of 1-9, which keeps measurement within the same order of magnitude. A preference of 1 indicates equality between two sub-indicators while a preference of 9 indicates that one sub-indicator is 9 times larger or more important than the one to which it is being compared. In this way comparisons are being made between pairs of sub-indicators where perception is sensitive enough to make a distinction. These comparisons result in a comparison matrix A (see an example in Table 6.4, for the TAI dataset) where $A_{ii} = 1$ and $A_{ij} = 1 / A_{ji}$.

Table 6.4. *Comparison matrix A of eight sub-indicators (semantic scale)*

Objective	Patents	Royalties	Internet	Tech exports	Telephones	Electricity	Schooling	University
Patents	1	2	3	2	5	5	1	3
Royalties	1/2	1	2	1/2	4	4	1/2	3
Internet	1/3	1/2	1	1/4	2	2	1/5	1/2
Tech. exports	1/2	2	4	1	4	4	1/2	3
Telephones	1/5	1/4	1/2	1/4	1	1	1/5	1/2
Electricity	1/5	1/4	1/2	1/4	1	1	1/5	1/2
Schooling	1	2	5	2	5	5	1	4
University	1/3	1/3	2	1/3	2	2	1/4	1

For the example shown in Table 6.4, Patents is three times more important than Internet, and consequently Internet has one-third the importance of Patents. Each judgment reflects, in reality, the perception of the ratio of the relative contributions (weights) of the two sub-indicators to the overall objective being assessed as shown in Table 6.5 for the first three sub-indicators.

Table 6.5. *Comparison matrix A for three sub-indicators*

Objective	Patents	Royalties	Internet	……
Patents	w_P/w_P	w_P/w_{ROY}	w_P/w_I	
Royalties	w_{ROY}/w_P	w_{ROY}/w_{ROY}	w_{ROY}/w_I	
Internet	w_I/w_P	w_I/w_{ROY}	w_I/w_I	
……				

The relative weights of the sub-indicators are calculated using an eigenvector technique. One of the advantages of this method is that it is able to check the consistency of the comparison matrix through the calculation of the eigenvalues. Figure 6.2 shows the results of the evaluation process

and the weights obtained (together with the corresponding standard deviation). The exercise was carried out at JRC interviewing experts in the field.

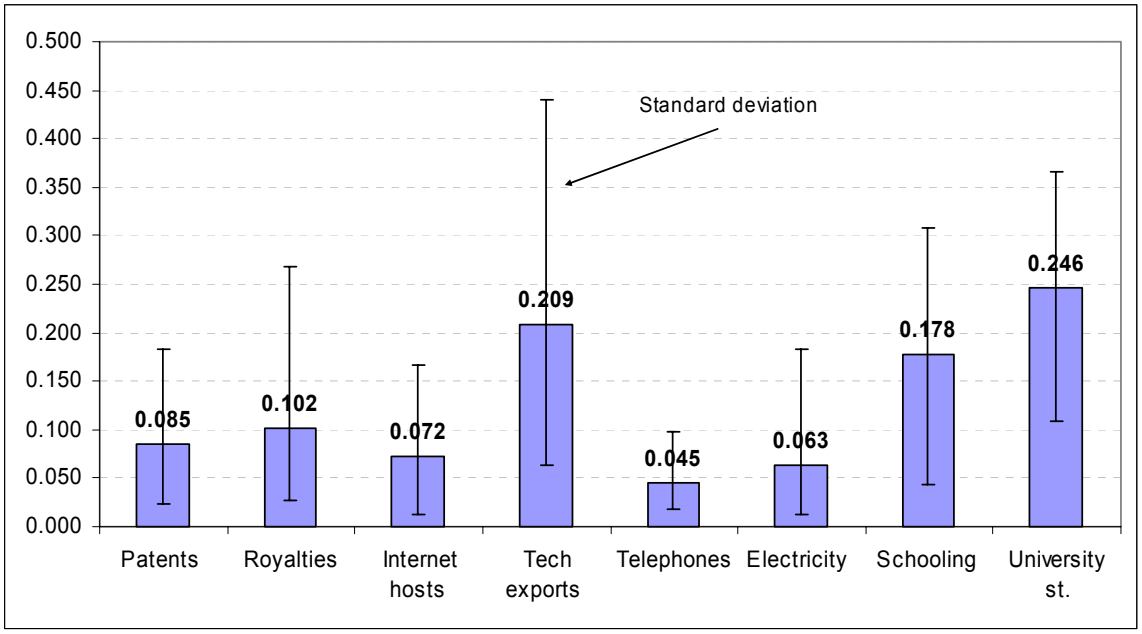


Figure 6.2. *Results of the AHP for the TAI example. Average weight (bold) and standard deviation.*

It is often the case that people's thinking is not always consistent. For example, if one claims that A is much more important than B, B slightly more important than C, and C slightly more important than A, judgment is inconsistent and decisions made are less trustworthy. Inconsistency, however, is part of the human nature and therefore in reality it is enough just to measure somehow the degree of inconsistency. This appears to be the only way so results could be defended and justified in front of public. AHP tolerates inconsistency through the amount of redundancy. For a matrix of size Q × Q only Q–1 comparisons are required to establish weights for Q indicators. The actual number of comparisons performed in AHP is Q(Q–1)/2. This redundancy has two opposite consequences: on one hand it is computationally costly, but on the other hand it is a useful feature as it is analogous to estimating a number by calculating the average of repeated observations. This results in a set of weights that are less sensitive to errors of judgment. In addition, this redundancy allows for a measure of these judgment errors by providing a means of calculating an inconsistency ratio (*Saaty, 1980; Karlsson, 1998*). According to Saaty small inconsistency ratios (less than 0.1 is the suggested rule-of-thumb, although even 0.2 is often cited) do no drastically affect the weights.

Analytic Hierarchy Process

Advantages	**Disadvantages**
▪ The method can be used both for qualitative and quantitative data ▪ The method increase the transparency of the composite	▪ The method requires a high number of pairwise comparisons and thus it can be computationally costly. ▪ The results depends on the set of evaluators chosen and the setting of the experiment

Examples of use

- Index of Environmental Friendliness, (Puolamaa et al., 1996)

6.1.9 Conjoint analysis

Merely asking respondents how much importance they attach to a sub-indicator is unlikely to yield effective "willingness to pay" valuations. Those can be inferred by using conjoint analysis from respondents' ranking of alternative scenarios (Hair et al. 1995). Thus weights equal willingness to pay.

The conjoint analysis (CA) is a decompositional multivariate data analysis technique frequently used in marketing (see McDaniel and Gates, 1998) and consumer research (see Green and Srinivasan, 1978). If AHP derives the "worth" of an alternative *summing up* the "worth" of the individual sub-indicators, the CA does the opposite, i.e. it disaggregates preferences. This method asks for an evaluation (a preference) over a set of alternative scenarios (a scenario can be thought as a given set of values for the sub-indicators). Then this preference is decomposed by relating the single components (the known values of sub-indicators of that scenario) to the evaluation.

Although this methodology uses statistical analysis to treat data, it operates with people (experts, politicians, citizens) who are asked to choose which set of sub-indicators they prefer, with each person presented with several different choice sets to evaluate. The absolute value (or level) of sub-indicators would be varied both within the choice sets presented to the same individual and across individuals. A preference function would be estimated using the information coming from the different scenarios. Therefore a probability of the preference could be estimated as a function of the levels of the sub-indicators defining the alternative scenarios:

$$pref_c = P(I_{1c}, I_{2c}, ..., I_{Qc}) \tag{6.10}$$

where I_{qc} is the level of sub-indicator $q=1,...,Q,$ for country $c=1,...,M$. After estimating this probability (often using discrete choice models), the derivatives with respect to the sub-indicators of the preference function can be used as weights to aggregate the sub-indicators in a composite index:

$$CI_c = \sum_{q=1}^{Q} \frac{\partial P}{\partial I_{qc}} I_{qc} \tag{11}$$

The idea is to calculate the total differential of the function P at the point of indifference between alternative states of nature. Solving for the sub-indicator q one obtains the marginal rate of substitution of I_{qc}. Therefore $\partial P / \partial I_{qc}$ (thus the weight) indicates a trade-off: how the preference changes with the change of the indicator. This implies the compensability among indicators, i.e. the possibility of offsetting the lack in some dimension with an outstanding performance in another dimension. As for other approaches already described, this is an important feature of this method and should be carefully evaluated vis à vis the objectives of the whole analysis (e.g. compensability might not be desirable when dealing with environmental issues).

Conjoint Analysis

Advantages	Disadvantages
▪ It obtains weights with the meaning of trade-offs ▪ It takes into account the socio-political context, and the values of respondents.	▪ It needs a pre-specified utility function and it implies compensability. ▪ Depends on the sample of respondent chosen and on how questions are framed ▪ It requires a large sample of respondents and each respondent may be required to express a large number of preferences. ▪ The estimation process is complex.

Examples of use

Indicator of quality of life in the city of Istanbul (Ülengin et al. 2001)
Advocated by Kahn (1998) and Kahn and Maynard (1996) for environmental applications.

6.1.10 Performance of the different weighting methods

To supply the reader with an idea of the diversity in weights obtained by applying the different methods we calculate the weights for the TAI example using four weighting methods (Table 6.6): EW, FA, BAL, AHP. Clearly with each method different sub-indicators are evaluated in a very different way. Patents, for example, are worth 17% of the weight according to the FA but only 9% according to the AHP. This deeply influences the variability of each country's ranking, as shown by Table 6.7 (BOD added). For examples the Republic of Korea ranks second with the AHP but only 5th when EW or FA are used. This is because AHP rewards with high weights (more than 20%) two indicators, *High tech exports* and *University enrolment ratio*, where this country has higher scores for one or both indicators as compared with USA, Sweden or Japan.

The role of the variability in the weights and their influence in the value of the composite will be the object of the section on sensitivity analysis (section 7).

Table 6.6. *Weights for the sub-indicators obtained using 4 different methods: equal weighting (EW), factor analysis (FA), budget allocation (BAL), and analytic hierarchy process (AHP)*

	Patents	*Royalties*	*Internet*	*Tech exports*	*Telephones*	*Electricity*	*Schooling*	*University*
EW	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
FA	0.17	0.15	0.11	0.06	0.08	0.13	0.13	0.17
BAL	0.11	0.11	0.11	0.18	0.10	0.06	0.15	0.18
AHP	0.09	0.10	0.07	0.21	0.05	0.06	0.18	0.25

Table 6.7. *Countries rank according to five different weighting methods*[*]

	EW	*FA*	*BOD*	*BAL*	*AHP*
Finland	1	1	1	1	1
United States	2	2	1	2	3
Sweden	3	3	1	3	4
Japan	4	4	4	5	5
Korea, Rep. of	5	5	6	4	2
Netherlands	6	6	7	8	11

United Kingdom	7	8	9	7	7
Singapore	8	11	12	6	6
Canada	9	10	8	10	10
Australia	10	7	10	11	9
Germany	11	12	11	9	8
Norway	12	9	5	13	16
Ireland	13	14	13	12	12
Belgium	14	15	15	14	13
New Zealand	15	13	14	17	18
Austria	16	16	16	15	15
France	17	17	17	16	14
Israel	18	18	18	18	17
Spain	19	19	20	19	19
Italy	20	20	19	21	21
Czech Republic	21	21	21	22	22
Hungary	22	23	23	20	20
Slovenia	23	22	22	23	23

(*) For example USA ranks first according to BOD, second according to EW, FA, and BAL and third according to AHP.

6.2 Aggregation techniques

The literature of composite indicators offers several examples of aggregation techniques. The most used are additive techniques that range from summing up country ranking in each sub-indicator to aggregating weighted transformations of the original sub-indicators. However, additive aggregations imply requirements and properties, both of component sub-indicators and of the associated weights, which are often not desirable, at times difficult to meet or burdensome to verify. To overcome these difficulties the literature proposes other, and less widespread, aggregation methods like multiplicative (or geometric) aggregations or non linear aggregations like the multi-criteria or the cluster analysis (the latter is explained in Section 3). This section reviews the most significant ones.

6.2.1 Additive methods

The simplest additive aggregation method entails the calculation of the ranking of each country according to each sub-indicator and the summation of resulting ranking (e.g. Information and Communication Technologies Index - Fagerberg J. 2001).

$$CI_c = \sum_{q=1}^{Q} Rank_{qc} \quad \text{for } c=1,...,M. \tag{6.12}$$

This method is therefore based on ordinal information. Its advantages are simplicity and the independence to outliers. Its disadvantage is that the method loses the absolute value information.

The second method proposed is based on the number of indicators that are above and below some benchmark. This method uses nominal scores for each indicator to calculate the difference between the number of indicators that are above and below an arbitrarily defined threshold around the mean (e.g.the Innovation Scoreboard of European Commission, 2001a).

$$CI_c = \sum_{q=1}^{Q} \cdot sgn\left[\frac{I_{qc}}{I_{EUq}} - (1+p)\right] \quad \text{for } c=1,...,M. \tag{6.13}$$

The threshold value p can be arbitrarily chosen above or below the mean. Pros and cons are the same as with the preceding method. Its advantages are simplicity and the fact that it is unaffected by outliers. The disadvantage is that it loses interval level information. For example, assume that the value of indicator I for country a is 30% above the mean and the value for country b is 25% above the mean, with a threshold of 20% above the mean. Both country a and b are then counted equally as 'above average', in spite of a having a higher score than b.

By far the most widespread linear aggregation is the summation of weighted and normalized sub-indicators:

$$CI_c = \sum_{q=1}^{Q} w_q I_{qc} \tag{6.14}$$

with $\sum_q w_q = 1$ and $0 \le w_q \le 1$, for all $q=1,..,Q$ and $c=1,...,M$.

Although widely used, this aggregation entails restrictions on the nature of sub-indicators. In order to obtain composite indexes fully and only reflecting the information contained in their components it is necessary to know the condition under which the weighted summation can be properly done. In particular the possibility to obtain a meaningful composite indicator depends upon conditions on the sub-indicators used for the aggregation and on the unit of measurement of these sub-indexes. Furthermore, additive aggregations have important implications, which are discussed next, as far the interpretation of weights is concerned.

6.2.2 Preference independence

When using a linear additive aggregation technique a necessary and sufficient condition for the existence of a proper composite indicator is preference independence: *given the sub-indicators* $\{x_1, x_2, ..., x_Q\}$*, an additive aggregation function exists if and only if these indicators are mutually preferentially independent*[22] (Debreu, 1960; Keeney and Raiffa, 1976; Krantz et al., 1971).
Preferential independence is a very strong condition since it implies that the trade-off ratio between two variables $S_{x,y}$ is independent of the values of the *Q-2* other variables, (Ting, 1971)[23]. From an *operational point of view* this means that an additive aggregation function permits the assessment of the marginal contribution of each variable separately. These marginal contribution can then be added together to yield a total value. If, for example, environmental dimensions are involved, the use of a linear aggregation procedure implies that among the different aspects of an ecosystem there are not phenomena of synergy or conflict. This appears to be quite an unrealistic assumption (Funtowicz et al., 1990). For example, *"laboratory experiments made clear that the combined impact of the acidifying substances SO_2, NO_X, NH_3 and O_3 on plant growth is substantially more severe that the (linear) addition of the impacts of each of these substances alone would be."* (Dietz and van der Straaten, 1992)".

What happens if the linear aggregation is nevertheless done? The resulting composite indicator will be biased, i.e. it will not entirely reflect the information of its sub-indicators. The dimension and the direction of the error are not easily determined, thus the correction of the composite can not be properly done.

---

[22] A subset of indicators *Y* is *preferentially independent* of Y^C (the complement of *Y*) only if any conditional preference among elements of *Y*, holding all elements of Y^C fixed, remain the same, regardless of the levels at which Y^C are held. The variables $x_1, x_2, ..., x_Q$ are *mutually preferentially independent* if every subset *Y* of these variables is preferentially independent of its complementary set of evaluators.

[23] $\frac{\partial S_{x,y}}{\partial z} = 0$, $\forall x, y \in Y, \forall z \in Y^C$, see the previous note above.

6.2.3 Weights and aggregations: lessons from multi-criteria analysis

The common practice in constructing composite indicators is well synthesized in a recent OECD report: "Greater weight should be given to components which are considered to be more significant in the context of the particular composite indicator" (OECDb, 2003, p. 10).

Yet, it can be proven that when using an additive or a multiplicative aggregation rule and sub-indicators are expressed as quantities (and not qualities) the substitution rates equal the weights of the variables up to a multiplicative coefficient[24] (Munda and Nardo 2003). As a consequence, **weights in additive aggregations necessarily have the meaning of substitution rates and do not indicate the importance of the indicator associated**. This implies a compensatory logic. Compensability refers to the existence of trade-offs, i.e. the possibility of offsetting a disadvantage on some variables by a sufficiently large advantage on other variables. For example, in the construction of the TAI index a compensatory logic (using equal weighting) would imply that one is willing to renounce, let's say, to 2% of *Patents granted to residents* or to 2% of *University enrolment* in exchange of a 2% increase in *Electricity consumption*.

The implication is the existence of a theoretical inconsistency in the way weights are actually used and their real theoretical meaning. For the weights to be interpreted as "importance coefficients" (in jargon symmetrical importance of variables, e.g. place the greatest weight beside the most important "dimension") non-compensatory aggregation procedures must be used to construct composite indicators (Podinovskii, 1994). This can be done using a non-compensatory multi-criteria approach.

A non-compensatory Multicriteria Approach (MCA)

The basic features of non compensatory multi-criteria are two: countries are ordered into binary relations in order to allow pair-wise comparisons, and the relationships created by the binary relations are *exploited* somehow. According to the way in which those two steps are performed several methods are available (see Munda 1995 for a review). One of these, suited to indicators, is presented next.

When various variables are used to evaluate, i.e. rank, a set of countries, some of these variables may be in favour of one country while other variables may be in favour of another. As a consequence a conflict among the variables exists. This conflict can be treated at the light of a non-compensatory logic and taking into account the absence of preferential independence within a discrete multi-criteria approach (Munda, 1995; Roy, 1996; Vincke, 1992).

Given a set of sub-indicators $\boldsymbol{G}=\{x_q\}$, $q=1,...,Q$, and a finite set $\boldsymbol{M}=\{c\}$, $c=1,...,M$ of countries, let's assume that the evaluation of each country c with respect to an individual indicator x_q (i.e. the

[24] Suppose that country *a* is evaluated according to some criteria/sub-indicators $(x_1(a),...,x_Q(a))$, then the *substitution rate at a*, of sub-indicator *j* with respect to sub-indicator *r* (taken as a reference) is the amount $S_{jr}(a)$ such that, country *b* whose evaluations are: $x_l(a)=x_l(b), \forall l \neq j,r$; $x_j(b)=x_j(a)-1$; and $x_r(b)=x_r(a)+S_{jr}(a)$ is indifferent to country *a*. Therefore, $S_{jr}(a)$ is the amount which must be added to the reference sub-indicator in order to compensate the loss of one unit on sub-indicator *j* keeping constant the others. While for additive aggregations the substitution rate is constant, in the multiplicative aggregation it is proportional to the relative score of the indicator with respect to the others.

indicator score or variable) is based on an *interval or ratio* scale of measurement. For simplicity of exposition, we assume that a higher value of an individual indicator is preferred to a lower one (the higher, the better). Let's also assume the existence of a set of weights $w=\{w_q\}$, $q=1,2,...Q$, with $\sum_{q=1}^{Q} w_q = 1$, interpreted as *importance coefficients*. This information constitutes the impact matrix. For explanatory purposes suppose to consider only 5 of the countries included in the TAI dataset[25] and give equal weighting to all the sub-indicators (Table 6.8):

Table 6.8. *Impact matrix for the TAI example, reduced dataset.*

	Patents	***Royalties***	***Internet***	***Tech exports***	***Telephones***	***Electricity***	***Schooling***	***University***
Finland	187	125.6	200.2	50.7	3.080	4.150	10	27.4
USA	289	130	179.1	66.2	2.997	4.073	12	13.9
Sweden	271	156.6	125.8	59.7	3.096	4.145	11.4	15.3
Japan	994	64.6	49	80.8	3.003	3.865	9.5	10
Korea	779	9.8	4.8	66.7	2.972	3.653	10.8	23.2
weight	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

The mathematical problem is then how to use this information to rank in a complete pre-order (i.e. without any incomparability relation) all the countries from the best to the worst one. Especially four points are important:

1. Intensity of preference (how much country *a* is better than country *b* according to sub-indicator *q*);
2. Number of indicators in favour of a given country;
3. Weight attached to each indicator;
4. Relationship of each country with respect to all the others.

The sources of uncertainty and imprecise assessment should be reduced as much as possible. Unfortunately Arrow's impossibility theorem (Arrow, 1963) clearly shows that no perfect aggregation convention can exist. Therefore, when aggregating it is essential to check not only which properties are respected by a given ranking procedure, but also that none of the essential properties for the specific problem faced is lost.

The mathematical aggregation convention can be divided into two main steps:

1. Pair-wise comparison of countries according to the whole set of sub-indicators used.
2. Ranking of countries in a complete pre-order.

The result of the first step is a $M \times M$ matrix, E, called *outranking matrix* (Arrow and Raynaud, 1986, Roy, 1996). Any generic element of E: e_{jk} , $j \neq k$ is the result of the pair-wise comparison, according to all the Q sub-indicators, between countries j and k. Such a global pair-wise comparison is obtained by means of equation:

[25] Data are not normalized. Normalization does not change the result of the multicriteria method whenever it does not change the ordinal information of the data matrix.

$$e_{jk} = \sum_{q=1}^{Q} \left(w_q(Pr_{jk}) + \frac{1}{2} w_q(In_{jk}) \right) \qquad (6.15)$$

where $w_q(Pr_{jk})$ and $w_q(In_{jk})$ are the weights of sub-indicators presenting a preference and an indifference relation respectively.
In other words, in Equation 6.15 above the score of country *j* is the sum of the weights of sub-indicators for which this country does better than country *I,* as well as – if any – half of the weights for the sub-indicators according to which the two countries do equally well.

It clearly holds $e_{jk} + e_{kj} = 1$. The pair-wise comparisons are different from those in the AHP method: there the question to be answered was whether I_q is more important than I_z, here, instead, the question is whether I_q is higher for country *a* or for country *b*. And if I_q is indeed higher for country *a* , it is the weight of sub-indicator *q* which enters into the computation of the overall importance of country *a*, in a way which is consistent with the definition of weights as importance measures.

In our example the pair-wise comparison of e.g. Finland and USA shows that Finland has better scores for the sub-indicators Internet (weight 1/8), Telephones (weight 1/8), Electricity (weight 1/8) and University (weight 1/8). Thus the score for Finland is 4*1/8=0.5 while the complement to one is the score of USA. The resulting outranking matrix is in Table 6.9:

Table 6.9. *Outranking matrix in the multicriteria analysis*

	Finland	***USA***	***Sweden***	***Japan***	***Korea***
Finland	0	0.5	0.375	0.75	0.625
USA	0.5	0	0.5	0.625	0.625
Sweden	0.625	0.5	0	0.75	0.625
Japan	0.25	0.375	0.25	0	0.75
Korea	0.375	0.375	0.375	0.25	0

The way in which these information are combined generates several possible ranking procedures (see Young 1988 and Munda 2004), each with pros and cons. One possible algorithm is the Condorcet-Kemeny-Young-Levenglick (CKYL) ranking procedure (Munda and Nardo 2003). According to CKYL the ranking of countries with the highest likelihood is the one supported by the maximum number of sub-indicators for each pair-wise comparison, summed over all pairs of countries considered. More formally, all the *M(M–1)* pair-wise comparisons compose the outranking matrix *E*. Call *R* the set of all *M!* possible complete rankings of alternatives, $R=\{r_S\}$, $s=1,2,...,M!$ For each r_S, compute the corresponding score φ_s as the summation of e_{jk} over all the $\binom{M}{2}$ pairs *j,k* of alternatives, i.e. $\varphi_s = \sum e_{jk}$ *where* $j \neq k,\ s = 1,2,...M!$ *and* $e_{jk} \in r_s$

The final ranking ($r*$) is the solution of:

$$r* \Leftrightarrow \varphi_* = \max \sum e_{jk} \qquad \text{where } e_{jk} \in R \qquad (6.16)$$

In our example the number of permutations obtained from 5 countries are 120, the first 5 are listed in Table 6.10. The score of, for example, the first ranking (USA, Sweden, Finland, Japan and Korea) is obtained as follows: according to the impact matrix the comparison of USA with the other countries yields 0.5 against Finland and Sweden, and 0.625 against Japan and Korea (overall 2.25). The comparison of Sweden yields 0.625 against Finland and Korea and 0.75 against Japan (overall 2). Finland obtains 0.625 against Korea and 0.75 against Japan (overall 1.375). Finally Japan obtains 0.75 against Korea. The final score of this ranking is then equal to 2.25+2+1.375+0.75=6.375.

Table 6.10. *Permutations obtained from the outranking matrix and associated score.*

USA	Sweden	Finland	Japan	Korea	**6.375**
Sweden	Finland	USA	Japan	Korea	**6.375**
Sweden	USA	Finland	Japan	Korea	**6.375**
Finland	USA	Sweden	Japan	Korea	6.125
Finland	Sweden	USA	Japan	Korea	6.125
USA	Finland	Sweden	Japan	Korea	6.125

According to expression (6.16) the final ranking will be the permutation(s) with the highest score. In our example the first 3 permutations have the highest overall score and thus all those can be considered as winning ranking.

This aggregation method has the advantage to overcome some of the problems raised by additive or multiplicative aggregations: preference dependence, the use of different ratio or interval scale to express the same indicator and the meaning of trade-offs given to the weights. With this method, moreover, qualitative and quantitative information can be jointly treated. In addition, it does not need any manipulation or normalization to assure the comparability of sub-indicators.

The drawbacks, instead, include the dependence of irrelevant alternatives, i.e. the possible presence of cycles/rank reversal in which in the final ranking, country *a* is preferred to *b*, *b* is preferred to *c* but *c* is preferred to *a* (the same problem highlighted for AHP with indicators). Furthermore, information on intensity of preference of variables is never used: if one indicator for country *a* is much less than the same indicator for country *b* produces the same ranking as the case in which this difference is very small[26]. Notice that with this method the focal point is shifted to the determination of weights, which becomes crucial for the result. Examples of MCA include agro-ecological indicators (Girardin et al., 2000) and an indicator of quality of life of three towns near to Puerto Vallarta, Mexico (Massam, 2002) within a project on the effects of tourism on the quality of life of small communities near international tourist resorts.

6.2.4 Geometric aggregation

As shown above, an undesirable feature of additive aggregations is the full compensability they imply: poor performance in some indicators can be compensated by sufficiently high values of other indicators. For example if a hypothetical composite were formed by inequality, environmental degradation, GDP per capita and unemployment, two countries, one with values 21, 1, 1, 1; and the other with 6,6,6,6 would have equal composite if the aggregation is additive.

[26] To obviate to this problem it is possible to set thresholds of this type: if the difference between two countries in the indicator I is more than x%, then give to the country with the highest score a much higher weight. If the difference is less than x% give nearly the same weight. However, more preciseness comes at the expenses of ad hoc threshold and weighting values.

Obviously the two countries would represent very different social conditions that would not be reflected in the composite. If multicriteria analysis entails full non-compensability, the use of a geometric aggregation (also called deprivational index) $CI_c = \prod_{q=1}^{Q} x_{q,c}^{w_q}$ is a in-between solution.

In our simple example the first country would have a much lower composite than the second if the aggregation is geometric (2.14 for the first and 6 for the second). The use of geometric aggregations can also be justified on the ground of the different incentives they supply to countries in a benchmarking exercise. Countries with low scores in some sub-indicators would prefer a linear rather than a geometric aggregation (the simple example above shows why). On the other hand the marginal utility from an increase in low absolute score would be much higher than in a high absolute score under geometric aggregation: the first country increasing by 1 unit the second indicator would increase its composite from 2.14 to 2.54, while country 2 would go from 6 to 6.23. In other terms the first country would increase its composite by 19% while the second only by 4%. The lesson is that a country should be more interested in increasing those sectors/activities/alternatives with the lowest score in order to have the highest chance to improve its position in the ranking if the aggregation is geometric rather than linear (Zimmermann and Zysno, 1983).

Furthermore the type of aggregation employed is strongly related with the method used to normalize raw data (Section 5). In particular Ebert and Welsch (2004) prove that the use of linear aggregations yields meaningful composite indicators only if data are all expressed in partially comparable interval scale (i.e. temperature in Celsius of Fahrenheit) of type $f: x \to \alpha x + \beta_i \quad \alpha > 0$ (i.e. α fixed, but β_i varying across subindicators) or in a fully comparable interval scale (β constant);
Non-comparable data measured in ratio scale (i.e. kilograms and pounds) $f: x \to \alpha_i x$ where $\alpha_i > 0$ (i.e. α_i varying across subindicators) can only be meaningfully aggregated by using geometric functions, provided that *x* is strictly positive. In other terms, except in the case of all indicators measured in different ratio scale, the measurement scale must be the same for all indicators when aggregating, thus care should be used when in the same composite coexist indicators measured in different scale: the normalization method used should properly remove the scale effect.

6.3 Conclusions: when to use what?

When using a model or an algorithm to describe a real-world issue formal coherence is a necessary property. Yet, formal coherence is not sufficient. The model in fact should fit objectives and intentions of the user, i.e. it must be the most appropriate tool for expressing the set of objectives that motivated the whole exercise. As explained in Section 2 the choice of which sub-indicators to use, how those are divided into classes, whether a normalization method has to be used (and which one), the choice of the weighting method, and how information is aggregated, all these features stem from a certain perspective on the issue to be modelled.

Table 6.11 highlights is the dependence of rankings to the aggregation methods used (in this case linear, geometric and based on the multicriteria technique for the TAI dataset with 23 countries). Although in all cases we used equal weighting, rankings produced are very different. For example Finland ranks first according to the linear aggregation, second according to the geometric aggregation and third according to the multicriteria. Notice that Korea ranks 16th with GME while is much above according to the other two methods, while the reverse happens for Belgium.

Table 6.11 *Rankings obtained using the linear and the geometric aggregations (resp. LIN and GME) and the multicriteria evaluation method (MCA). Dataset TAI, for 23 countries. Numbers refer to the position in ranking.*

	LIN	MCA	GME
Finland	1	3	2
United States	2	1	1
Sweden	3	2	3
Japan	4	4	4
Korea, Rep. of	5	9	16
Netherlands	6	8	5
United Kingdom	7	5	6
Singapore	8	12	18
Canada	9	11	13
Australia	10	9	14
Germany	11	7	8
Norway	12	6	11
Ireland	13	13	7
Belgium	14	17	9
New Zealand	15	15	17
Austria	16	15	12
France	17	14	10
Israel	18	18	15
Spain	19	20	19
Italy	20	19	21
Czech Republic	21	21	23
Hungary	22	23	22
Slovenia	23	22	20

The absence of an "objective" way of constructing composites should not result in a rejection of whatever type of composite. Composites can meaningfully supply information provided that the relation between the framing of a problem and the outcome in the the decision space are made

clear. A backward induction exercise could be useful in this context. Once the context and the modeller's objectives have been made explicit, the user can verify whether and how the selected model fulfils those objectives. A plurality of methods (all with their implications) can in principle be used and no model is a priori better than another, provided internal coherence is assured. In practice, different models can meet different expectations and stakes. Therefore, stakes must be made clear, and transparency should guide the entire process.

With this in mind we present a number of considerations that should help the reader in choosing the appropriate weighting and aggregation method. Table 6.12 presents the feasible combinations between the various aggregation and weighting methods.

Weighting methods

(a) The ***equal weighting*** can be applied after a proper scaling of the sub-indicators. Equal weighting works well if all dimensions (economic, social, environmental, etc.) are represented in the composite with the same number of sub-indicators (as in the TAI example). If this does not happen equal weighting implies a higher weight to the dimension represented by the larger number of components. Equal weighting is also appealing when high correlation of components indicators does not mean redundancy of information in the composite, i.e. when correlated components explain different aspects of the picture the composite aims to capture.

(b) ***Principal Components Analysis*** is a very useful exploratory technique to examine the correlation structure of groups of variables, and ***Factor Analysis*** is usually employed as a supplementary method with a view to examine thoroughly the relationships among the sub-indicators. However, there are two crucial problems with these arguments. First, weights assigned to sub-indicators in both of these techniques are based on correlations which do not necessarily correspond to the underlying relationships between the sub indicators and the phenomena being measured. In other words there is confusion between correlation and redundancy: redundancy implies correlation but the reverse is not necessarily true. Secondly, being based only on correlation, FA is a way to discipline homogeneity rather than to represent plurality. FA in fact can only be applied when variables are correlated, i.e. when they move in the same direction (if the correlation is positive, and in opposite direction if the correlation is negative).

(c) The ***Benefit of the Doubt*** approach is extremely parsimonious as regards the weighting assumptions, because it lets the data decide on the weighting issue, and it is sensible to national priorities. It is argued, though, that weights are country specific and that there are a number of technical problems involved in the estimation. Furthermore, the BOD method over-scores outliers.

(d) ***Multiple regression models*** can handle a large number of indicators. This approach can be applied in cases where the sub-indicators considered as input to the model are related to various policy actions and the output of the model is the target. The regression model, thereafter, could quantify the relative effect of each policy action on the output, i.e. the single indicator. However, this implies the existence of a "dependent variable" (not in the form of a composite indicator) that accurately and satisfactorily measures the target in question. Measuring the influence of a number of independent variables on this policy target is a reasonable question. Alternatively such an approach could be used for forecasting purposes.

In a more general case of multiple output indicators, *canonical correlation analysis* that is a generalization of multiple regression could be applied. However, in any case, there is always the uncertainty that the relations, captured by the regression model for a given range of inputs and output, may not be valid for different ranges.

(e) ***Unobserved components*** is similar in spirit to the multiple regression models, it does not need an explicit value for the "dependent variable" as it treats it like another unknown variable to estimate. This advantage is counter-balanced by the inconvenient of the complexity in estimation and the computational expensiveness.

(f) ***Participatory methods*** constitute a way to involve experts, citizens or politicians in the issue. Using policy goals as targets convinces the policy makers of the "soundness" of the weighting method, as long as those policy makers have defined the policy targets themselves. This approach is technically feasible when there is a well-defined basis for a certain policy, such as a National Policy Plan or similar reference documents. For international comparisons, such references are often not available, or they deliver contradictory results. Another counter-argument for the use of policy goals as targets is that the benefits of a given policy must be valued independently of the existing policy goals.

(g) ***Expert judgement*** is adopted when it is essential to bring together experts that have a wide spectrum of knowledge, experience and concerns, so as to ensure that a proper weighting system is found for a given application. The ***budget allocation*** is optimal for a maximum number of 10-12 indicators. If a too large number of indicators is involved, this method can give serious cognitive stress to the experts who are asked to allocate the budget.

(h) ***Public opinion polls*** have been extensively employed for many years for the setting of weights. In public opinion polls, issues are selected which are already on the public agenda, and thus receive roughly the same attention by the media. In many case studies, public opinion polls in different countries and years resulted in similar weighting schemes for certain environmental problems, which indicates that public opinion about the main threats to the environment is remarkably stable across both space and time. Therefore fears that the public evaluates environmental issues on an irrational basis, and therefore weights base upon public opinion will produce instability, appear to be unfounded.

(i) The ***Analytic Hierarchy Process*** is a widely used technique for multi-attribute decision making, as weighting method enables the decision-maker to derive weights as opposed to arbitrarily assigning them. An advantage of AHP is that unlike many other methods based on Utility Theory, its use for the purposes of comparisons does not require a universal scale. Furthermore, AHP tolerates inconsistency in the way people think through the amount of redundancy (more equations are available than the number of weights to be defined). This redundancy is a useful feature as it is analogous to estimating a number by calculating the average of repeated observations. The resulting weights are less sensitive to errors of judgments. These advantages may render the weights derived from AHP defensible d and justifiable in front of the public.

(j) ***Conjoint analysis*** derives the worth of the single sub-indicators from the worth of a composite, i.e. it reverses the process of AHP, with which it shares advantages and disadvantages. Further complication is the need to specify and estimate and utility function.

Aggregation methods

Linear aggregation method is useful when all sub-indicators have the same measurement unit and further ambiguities due to the scale effects have been neutralized, while ***geometric aggregations*** are appropriate when non-comparable and strictly positive sub-indicators are expressed in different ratio-scales. The absence of synergy or conflict effects among the indicators is a necessary condition to admit either linear or geometric aggregations. Furthermore, linear aggregations reward base-indicators proportionally to the weights, while geometric aggregations reward more those countries with higher scores.

In both linear and geometric aggregations **weights express trade-offs** between indicators: the idea is that deficits in one dimension can be offset by surplus in another. With linear aggregations the compensability is constant, while with geometric aggregations compensability is lower when the composite contains indicators with low values. In policy terms if compensability is admitted (as in the case of pure economic indicators) a country with low scores on one indicator will need much higher score on the others to improve its situation if the aggregation of information is geometric. Thus in a benchmarking exercise countries with low scores should prefer a linear rather than a geometric aggregation. On the other hand the marginal utility of an increase in the score would be much higher when the absolute value of the score is low. The resulting lesson is that a country should be more interested in increasing those sectors/activities/alternatives with the lowest score in order to have the highest chance to improve its position in the ranking if the aggregation is geometric. The opposite is true, i.e. a country has interest in specialising along its most effective dimensions, when the aggregation is linear.[27]

When different goals are equally legitimate and important, then a non compensatory logic may be necessary. This is usually the case when very different dimensions are involved in the composite, like in the case of environmental indexes, where physical, social and economic figures must be aggregated. If the analyst decides that an increase in economic performance can not compensate a loss in social cohesion or a worsening in environmental sustainability, then neither the linear nor the geometric aggregation are suitable. Instead, a non-compensatory ***multicriteria approach*** will assure non compensability by formalizing the idea of finding a compromise between two or more legitimate goals.

Multicriteria analysis, like any other method, has pros and cons. At least in its basic form this approach does not reward outliers, i.e. those countries having large advantages (disadvantages) in sub-indicators since it keeps only the ordinal information. Another disadvantage is the computational expensiveness when the number of countries is high (the number of permutations to calculate grows exponentially).

Table 6.12. *When to use what: compatibility between aggregation and weighting methods.*

[27] Compensability of aggregations is widely studied in fuzzy sets theory, for example Zimmermann and Zysno (1983) use the geometric operator $(\prod_q I_q)^{(1-\gamma)}(1-\prod_q(1-I_q))^{\gamma}$ where γ is a parameter of compensation: the larger is γ the higher is the degree of compensation between operators (in our case sub-indicators).

Weighting methods	*Aggregation methods*		
	Linear[4]	Geometric[4]	Multi-criteria
EW	Yes	Yes	Yes
PCA/FA	Yes	Yes	Yes
BOD	Yes[1]	No[2]	No[2]
UCM	Yes	No[2]	No[2]
BAL	Yes	Yes	Yes
AHP	Yes	Yes	No[3]
CA	Yes	Yes	No[3]

1 normalized with the maximin method.
2 BOD requires additive aggregation, similar arguments apply to UCM
3 At least with the multi-criteria methods requiring weights as importance coefficients.
4 With both linear and geometric aggregations weights need to trade-offs and not "importance" coefficients

7. Uncertainty and sensitivity analysis

The reader will recall from the introduction that composite indicators may send misleading, non-robust policy messages if they are poorly constructed or misinterpreted. The cons also mentioned that the construction of composite indicators involves stages where judgement has to be made: the selection of sub-indicators, the choice of a conceptual model, the weighting of indicators, the treatment of missing values etc. All these sources of subjective judgement will affect the message brought by the CIs in a way that deserve analysis and corroboration. A combination of uncertainty and sensitivity analysis can help to gauge the robustness of the composite indicator, to increase its transparency and to help framing a debate around it.

General procedures to assess uncertainty in composite indicators building are described in this section.

In particular, we shall try to tackle all possible sources of uncertainty, which arise from:

i. selection of sub-indicators,
ii. data quality,
iii. data editing,
iv. data normalisation,
v. weighting scheme,
vi. weights' values,
vii. composite indicator formula

Two combined tools are suggested: Uncertainty Analysis (UA) and Sensitivity Analysis (SA). UA focuses on how uncertainty in the input factors propagates through the structure of the composite indicator and affects the composite indicator values. SA studies how much each individual source of uncertainty contributes to the output variance.

In the field of building composite indicators, UA is more often adopted than SA (Jamison and Sandbu, 2001; Freudenberg, 2003) and the two types of analysis are almost always treated separately. A synergistic use of UA and SA is proposed and presented here, considerably extending earlier attempts in this direction (Tarantola *et al.*, 2000). We will exemplify it with the TAI example by building an error propagation analysis as complete as possible given the example, to the effect of showing the UA, SA machinery at work on a rather complicate setting,

e.g. one where one wants to test different index architectures. In practical applications it might happen that the aggregation formula and the weighing scheme will be dictated by the purpose of the index and/or by an agreement among the parties involved in the index construction and use, thus making the UA/SA simpler.

With reference to the uncertainty sources (i to vii above), the approach taken to propagate uncertainties could include in theory all of the steps below:

i. inclusion – exclusion of sub-indicators,
ii. modelling of data error, e.g. based of available information on variance estimation.
iii. alternative editing schemes, e.g. multiple imputation, described in section 4.
iv. using alternative data normalisation schemes, such as rescaling, standardisation, use of raw data.
v. using several weighting schemes, i.e. two methods in the participatory family (budget allocation **BAL** and analytic hierarchy process **AHP**), and one based on endogenous weighting (benefit of the doubt **BOD**)
vi. using several aggregation systems, i.e. linear **LIN**, another based on geometric mean of un-scaled variable **GME** and finally one based on multi-criteria ordering **MCA**, all described in Section II-6 above.
vii. weights' values, sampled from distributions when appropriate to the weighting scheme.

First TAI analysis. In a first analysis, in order to use the geometric mean aggregation approach GME, we shall omit (iii), i.e. we shall discard all countries with incomplete information. This is because even with imputation, we might generate zeros that might be untreatable by GME. In a second analysis described later we shall relax this assumption. Also modelling of the data error, point (ii) above, will not be included as in the case of TAI no standard error estimate is available for the sub-indicators. In a general case, based on estimate of the standard error associated to each individual sub-indicator, we could sample an error for each assuming a Gaussian error distribution, e.g., sampling a random number in the $[0,1]$ interval and mapping it into the cumulative probability density function (cpdf) of the sub-indicator error.

Furthermore, not all combinations of choices under (i) to (vii) above are feasible with our TAI index. In particular (see also Table 6.12 on previous section):

A. When using LIN for aggregation and BAL or AHP for weighting, the option "use of raw data" for normalisation is forbidden.
B. When using LIN for aggregation and BOD for weighting, the options "use of raw data" and "standardisation" for normalisation are forbidden.
C. When using GME for aggregation, then BOD for weighting is forbidden. Furthermore when using BAL and AHP, the option "standardisation" for normalisation is forbidden.
D. When using MCA for aggregation, then BOD for weighting is forbidden.

A few technicalities are also worth mentioning.

E. As all weighs for both AHP and BAL are given by the experts, we sample the expert rather than the weight to preserve coherence among weights, e.g. to avoid generating combinations of weights that no expert would have advocated for.

F. When using BOD, the exclusion of an indicator leads to a total re-run of the optimisation algorithm. When using BAL or AHP a simple rescaling of the weights to unit is sufficient.

Second TAI analysis. This differs from the first analysis in that we assume that stakeholders have converged to using LIN aggregation. In this case we can allow for alternative editing schemes, point (iii) above and consider all countries as in the original TAI. This analysis aims at answering mainly two questions:

(a) Does the use of one strategy versus another in indicator building (i to vii above) provide a biased picture of the countries' performance? How does this compare to the original TAI index?

(b) To what extent do the uncertain input factors (used to generate the alternatives i to vii above) affect the countries' ranks with respect to the original, deterministic TAI?

7.1 Set up of the analysis

7.1.1 Output variables of interest

Let

$$CI_c = f_{rs}\left(I_{1,c}, I_{2,c}, ...I_{Q,c}, w_{s,1}, w_{s,2}, ...w_{s,Q}\right) \quad (7.1)$$

be the index value for country *c, c=1,...,M,* according the weighting model $f_{rs}, r = 1,2,3, \quad s = 1,2,3$ where the index *r* refers to the aggregation system (LIN, GME, MCA) and index *s* refers to the weighting scheme (BAL, AHP, BOD). The index is based on *Q* sub-indicators $I_{1,c}, I_{2,c}, ...I_{Q,c}$ for that country and scheme-dependent weights $w_{s,1}, w_{s,2}, ...w_{s,Q}$ for the sub-indicators.

The rank assigned by the composite indicator to a given country, i.e. $Rank(CI_c)$ will be an output of interest studied in our uncertainty – sensitivity analysis.

Additionally, the average shift in countries' rank will be explored. This latter statistic captures in a single number the relative shift in the position of the entire system of countries. It can be quantified as the average of the absolute differences in countries' rank with respect to a reference ranking over the *M* countries:

$$\overline{R}_S = \frac{1}{M}\sum_{c=1}^{M}\left|Rank_{ref}(CI_c) - Rank(CI_c)\right| \quad (7.2)$$

The reference ranking for the TAI analysis is the original rank given to the country by the original version of the index.

The investigation of $Rank(CI_c)$ and $\overline{R}_S$ will be the scope of the uncertainty and sensitivity analysis (both in the first and second TAI analysis), targeting the questions raised in the introduction on the quality of the composite indicator. We always work on $Rank(CI_c)$ rather than on the raw values of the index CI_c as the multi criteria approach MCA only produces ranks for countries, as explained in Section II-6 of the present guidelines.

7.1.2 General framework for the analysis

As described in the following sections, we shall frame the analysis as a single Monte Carlo experiment, e.g. by plugging all uncertainty sources simultaneously, as to capture all possible synergistic effects among uncertain input factors. This will involve the use of triggers, e.g. the use of uncertain input factors used to decide e.g. which aggregation system and weighting scheme to adopt. To stay with the example, a discrete uncertain factor which can take integer values between 1 and 3 will be used to decide upon the aggregation system and another also varying in the same range for the weighting scheme. Other trigger factors will be generated to select which indicators to omit, the editing scheme (for the second TAI analysis only), the normalisation scheme and so on, till a full set of input variables is available to compute for the given run the statistics $Rank(CI_c)$, $\overline{R}_S$ described above.

7.1.3 Inclusion – exclusion of individual sub- indicators

No more than one indicator at a time is excluded for simplicity. A single random variable is used to decide if any indicator will be omitted and which one. Note that an indicator can also be practically neglected as a result of the weight assignment procedure. Imagine a very low weight is assigned by an expert to a sub-indicator q. Every time we select that expert in a run of the Monte Carlo simulation, the relative sub-indicator q will be almost neglected for that run.

7.1.4 Data quality

This is not considered here as discussed above.

7.1.5 Normalisation

As described in Section II-5 several methods are available to normalise sub-indicators. The methods that are most frequently met in the literature are based on the re-scaled values (equation (7.3a)) or on the standardised values (equation (7.3b)) or on the raw indicator values (7.3c).

$$\begin{cases} I_{q,c} = \dfrac{x_{q,c} - \min(x_q)}{range(x_q)} & (7.3a) \\ I_{q,c} = \dfrac{x_{q,c} - mean(x_q)}{std(x_q)} & (7.3b) \\ I_{q,c} = x_{q,c} & (7.3c) \end{cases}$$

where $I_{q,c}$ is the normalised and $x_{q,c}$ is the raw value of the sub-indicator x_q for country c.

Equations (7.3a) will be used in conjunction with all weighting schemes (BAL, AHP and BOD) for all aggregation systems (LIN, GME, MCA). Equation (7.3b) will be used in conjunction with weighting schemes (BAL, AHP) for aggregation systems (LIN, MCA). Finally, Equation (7.3c) will be used in conjunction with weighting schemes (BAL, AHP) for aggregation systems (GME, MCA).

7.1.6 Uncertainty analysis

All points of the (i) to (vii) chain of composite indicator building can introduce uncertainty in the output variables $Rank(CI_c)$ and $\overline{R}_S$. Thus we shall translate all these uncertainties into a set of scalar input factors, to be sampled from their distributions. As a result, all outputs $Rank(CI_c)$ and $\overline{R}_S$ are non-linear functions of the uncertain input factors, and the estimation of the probability distribution functions (pdf) of $Rank(CI_c)$ and $\overline{R}_S$ is the purpose of the uncertainty analysis. The UA procedure is essentially based on simulations that are carried on the various equations that constitute our *model*. As the model is in fact a computer programme that implements steps (i) to (vii) above, the uncertainty analysis acts on a *computational model*. Various methods are available for evaluating output uncertainty.

In the following, the Monte Carlo approach is presented, which is based on performing multiple evaluations of the model with k randomly selected model input factors. The procedure involves six steps:

Step 1. Assign a pdf to each input factor X_i, $i = 1,2...k$. The first input factor, X_1 is used for the selection of the editing scheme (for the second TAI analysis only):

X_1	Editing
1	Use bivariate correlation to impute missing data
2	Assign zero to missing datum

The second input factor X_2 is the trigger to select the normalisation method.

X_2	Normalisation
1	Rescaling (Equation 7.3a)
2	Standardisation (Equation 7.3b)
3	None (Equation 7.3c)

Both X_1 and X_2 are discrete random variables. In practice, they are generated drawing a random number ζ uniformly distributed in [0,1] and applying the so called Russian roulette algorithm, e.g. for X_1 we select 1 if $\zeta \in [0,0.5)$ and 2 if $\zeta \in [0.5,1]$. Uncertain factor X_3 is generated to select which sub-indicator –if any, should be omitted. The procedure is

ζ	X_3, excluded sub-indicator
$[0, \frac{1}{Q+1})$	None ($X_3 = 0$) all subindicators are used)
$[\frac{1}{Q+1}, \frac{2}{Q+1})$	$X_3 = 1$
…	…
$[\frac{Q}{Q+1}, 1]$	$X_3 = Q$

i.e. with probability $\frac{1}{Q+1}$ no sub-indicator will be excluded, while with probability [1-$\frac{1}{Q+1}$] one of the Q sub-indicators will be excluded with equal probability. Clearly we could have made the probability of $X_3 = 0$ larger or smaller than $\frac{1}{Q+1}$ and still sample the values $X_3 = 1,2,...Q$ with equal probability. We anticipate here that a scatter-plot based sensitivity analysis will allow us to track which indicator – when excluded – affects the output the most. Also recall that whenever a sub-indicator is excluded, the weights of the other factors are rescaled to 1 to make the composite index comparable if either BAL or AHP is selected. When BOD is selected the exclusion of a sub-indicator leads to a re-execution of the optimisation algorithm.

Trigger X_4 is used to select the aggregation system

X_4	Scheme
1	**LIN**
2	**GME**
3	**MCA**

Note that when LIN is selected the composite indicators are computed as

$$CI_c = \sum_{q=1}^{Q} w_{sq} I_{q,c} \tag{7.4}$$

while when GME is selected they are:

$$CI_c = \prod_{q=1}^{Q} \left(I_{q,c}\right)^{w_{4,q}} \tag{7.5}$$

When MCA is selected the countries are ranked directly from the outscoring matrix as described in section 6.

X_5 is the trigger to select the weighting scheme;

X_5	Scheme
1	**BAL**
2	**AHP**
3	**BOD**

The last uncertain factor X_6 is used to select the expert. In our experiment we had 20 expert, and once an expert is selected at runtime via the trigger X_6, the weights assigned by that expert (either for the BAL or AHP schemes) are assigned to the data. Clearly the selection of the expert has no bearing when BOD is selected ($X_5 = 3$). All the same this uncertain factor will be generated at each individual Monte Carlo simulation. This is because the row dimension of the Monte Carlo sample (called constructive dimension) should be fixed in a Monte Carlo experiment, i.e. even if some of the sampled factors will not be active at a particular run, they will be all the same generated by the random sample generation algorithm.

The constructive dimension of this Monte Carlo experiment, e.g. the number of random numbers to be generated for each trial, is hence $k = 6$.

Note that alternative arrangements of the analysis would have been possible. We shall return on this point in our discussion of the results.

Step 2. Having generated the input factors distributions in step 1, we can now generate randomly N combinations of independent input factors $\mathbf{X}^l$, $l = 1,2,...N$ (a set $\mathbf{X}^l = X_1^l, X_2^l, \ldots, X_k^l$ of input factors is called a sample). For each trial sample $\mathbf{X}^l$ the computational model can be evaluated, generating values for the scalar output variable Y^l, where Y^l is either $Rank(CI_c)$, the value of the rank assigned by the composite indicator to each country, or $\overline{R}_S$, the averaged shift in countries' rank.

Step 3. We can now close the loop over *l*, and analyse the resulting output vector $\boldsymbol{Y}^l$, with *l* = 1, ..., *N*.

The generation of samples can be performed using various procedures, such as simple random sampling, stratified sampling, quasi-random sampling or others (Saltelli *et. al*, 2000a). The sequence of $\boldsymbol{Y}^l$ allows the empirical pdf of the output *Y* to be built. The characteristics of this pdf, such as the variance and higher order moments, can be estimated with an arbitrary level of precision that is related to the size of the simulation N.

7.1.7 Sensitivity analysis using variance-based techniques

A necessary step when designing a sensitivity analysis is to identify the output variables of interest. Ideally these should be relevant to the issue tackled by the model, as opposed to just relevant to the model *per se* (Saltelli *et al.*, 2000b, 2004).

In the following, we shall apply sensitivity analysis to output variables $Rank(CI_c)$, and $\overline{R}_S$, for their bearing on the quality assessment of our composite indicator.

It has been noted earlier in this work that composite indicators can be considered as models. When –as in the present analysis- several layers of uncertainty are simultaneously activated, composite indicators turn out to be non linear, possibly non additive models. As argued by practitioners (Saltelli et al., 2000a, EPA, 2004), robust, "model-free" techniques for sensitivity analysis should be used for non linear models. Variance-based techniques for sensitivity analysis are model free and display additional properties convenient for the present analysis:

- they allow an exploration of the whole range of variation of the input factors, instead of just sampling factors over a limited number of values, as done e.g. in fractional factorial design (Box *et al.* 1978);
- they are quantitative, and can distinguish main effects (first order) from interaction effects (higher order).
- they are easy to interpret and to explain

- they allow for a sensitivity analysis whereby uncertain input factors are treated in groups instead of individually
- they can be justified in terms of rigorous settings for sensitivity analysis, as we shall discuss later in this section.

How do we compute a variance based sensitivity measure for a given input factor X_i? We start from the fractional contribution to the model output variance (i.e. the variance of Y where Y is either $Rank(CI_c)$, and $\overline{R}_S$) due to the uncertainty in X_i. This is expressed as:

$$V_i = V_{X_i}(E_{\mathbf{X}_{-i}}(Y|X_i)) \tag{7.6}$$

One way of reading Equation (7.6) is the following. Imagine to fix factor X_i, e.g. to a specific value x_i^* in its range, and to compute the mean of the output Y averaging over all factor but factor X_i: $E_{\mathbf{X}_{-i}}(Y|X_i = x_i^*)$. Imagine then to take the variance of the resulting function of x_i^* over all possible x_i^* values. The result is given by Equation (7.6), where the dependence from x_i^* has been dropped since we have averaged over it. V_i is a number between 0 (when X_i does not gives a contribution to Y at the first order), and $V(Y)$, the unconditional variance of Y, when all factors other than X_i are non influent at any order. The meaning of "order" will be explained in a moment. Note that it is always true that:

$$V_{X_i}(E_{\mathbf{X}_{-i}}(Y|X_i)) + E_{X_i}(V_{\mathbf{X}_{-i}}(Y|X_i)) = V(Y) \tag{7.7}$$

where the first term in (7.7) is called a main effect, and the second one the residual. An important factor should have a small residual, e.g. a small value of $E_{X_i}(V_{\mathbf{X}_{-i}}(Y|X_i))$. This is intuitive as the residual measures the expected reduced variance that one would achieve if one could fix X_i. Let us write this as $V_{\mathbf{X}_{-i}}(Y|X_i = x_i^*)$, a variance conditional on x_i^*. Then the residual $E_{X_i}(V_{\mathbf{X}_{-i}}(Y|X_i))$ is the expected value of such conditional variance, averaged over all possible values of x_i^* and this should be small if is X_i influential. A first order sensitivity index is obtained by normalised the first order term by the unconditional variance:

$$S_i = \frac{V_{X_i}(E_{\mathbf{X}_{-i}}(Y|X_i))}{V(Y)} = \frac{V_i}{V(Y)} \tag{7.8}$$

One can compute conditional variances corresponding to more than one factors, e.g. for two factors X_i and X_j one can compute $V_{X_iX_j}(E_{\mathbf{X}_{-ij}}(Y|X_i, X_j))$, and from this a second order term variance contribution can be written as:

$$V_{ij} = V_{X_iX_j}(E_{\mathbf{X}_{-ij}}(Y|X_i, X_j)) - V_{X_i}(E_{\mathbf{X}_{-i}}(Y|X_i)) - V_{X_j}(E_{\mathbf{X}_{-j}}(Y|X_j)) \tag{7.9}$$

where clearly V_{ij} is only different from zero if $V_{X_iX_j}(E_{\mathbf{X}_{-ij}}(Y|X_i, X_j))$ is larger than the sum of the first order term relative to factors X_i and X_j.

When all *k* factors are independent from one another, the sensitivity indices can be computed using the following decomposition formula for the total output variance *V(Y)*

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \sum_i \sum_{j>i} \sum_{\substack{l>j \\ j>i}} V_{ijl} + \ldots + V_{12\ldots k} \tag{7.10}$$

Terms above the first order in equation (7.10) are known as interactions. A model without interactions among its input factors is said to be additive. In this case, $\sum_{i=1}^{k} V_i = V(Y)$, $\sum_{i=1}^{k} S_i = 1$ and the first order conditional variances of equation (7.6) are all what we need to know to decompose the model output variance. For a non-additive model, higher order sensitivity indices, responsible for interaction effects among sets of input factors, have to be computed. However, higher order sensitivity indices are usually not estimated, as in a model with *k* factors the total number of indices (including the S_i's) that should be estimated is as high as 2^k-1. For this reason, a more compact sensitivity measure is used. This is the total effect sensitivity index, which concentrates in one single term all the interactions involving a given factor X_i. To exemplify, for a model of *k*=3 independent factors, the three total sensitivity indices would be:

$$S_{T1} = \frac{V(Y) - V_{X_2X_3}(E_{X_1}(Y|X_2, X_3))}{V(Y)} = S_1 + S_{12} + S_{13} + S_{123} \tag{7.11}$$

And analogously:

$$S_{T2} = S_2 + S_{12} + S_{23} + S_{123} \tag{7.12}$$
$$S_{T3} = S_3 + S_{13} + S_{23} + S_{123}$$

The conditional variance $V_{X_2X_3}(E_{X_1}(Y|X_2, X_3))$ in equation (7.11) can be written in general terms as $V_{\mathbf{X}_{-i}}(E_{X_i}(Y|\mathbf{X}_{-i}))$ (Homma and Saltelli, 1996). It expresses the total contribution to the variance of *Y* due to non-X_i i.e. to the *k*-1 remaining factors, so that $V(Y) - V_{\mathbf{X}_{-i}}(E_{X_i}(Y|\mathbf{X}_{-i}))$ includes all terms, i.e. a first order as well as interactions in equation (7.10), that involve factor X_i. In general $\sum_{i=1}^{k} S_{Ti} \geq 1$.

Given the algebraic relation (7.7), the total effect sensitivity index can also be written as:

$$S_{Ti} = \frac{V(Y) - V_{\mathbf{X}_{-i}}(E_{X_i}(Y|\mathbf{X}_{-i}))}{V(Y)} = \frac{E_{\mathbf{X}_{-i}}(V_{X_i}(Y|\mathbf{X}_{-i}))}{V(Y)} \tag{7.13}$$

For a given factor X_i a significant difference between S_{Ti} and S_i flags an important role of interactions for that factor in *Y*. Highlighting interactions among input factors helps us improve our understanding of the model structure. Estimators for both (S_i, S_{Ti}) are provided by a variety of methods reviewed in Chan *et al.* (2000). Here the method of Sobol' (1993), in its improved

version due to Saltelli (2002) is used. The method of Sobol' uses quasi-random sampling of the input factors. The pair (S_i, S_{Ti}) give a fairly good description of the model sensitivities at a reasonable cost, which for the improved Sobol' method is of $2n(k+1)$ model evaluations, where n represents the sample size required to approximate the multidimensional integration implicit in the E and V operators above to a plain sum. n can vary in the hundred-to-thousand range.

When the uncertain input factors X_i are dependent, the output variance cannot be decomposed as in equation (10). The S_i, S_{Ti} indices, as defined by (7.6) and (7.13) are still valid sensitivity measures for X_i, though their interpretation changes as, e.g. S_i carries over also the effects of other factors that can be positively or negatively correlated to X_i (see Saltelli and Tarantola, 2002), while S_{Ti} can no longer be decomposed meaningfully into main effect and interaction effects. The usefulness of S_i, S_{Ti}, also for the case of non-independent input factors, is also linked to their interpretation in terms of "settings" for sensitivity analysis.

We offer here without proof a description of two settings linked to S_i, S_{Ti}. A justification is in Saltelli et al., 2004.

Factors' Prioritisation (FP) Setting. One must bet on a factor that, once "discovered" in its true value and fixed, would reduce the most V(Y). Of course one does not know where the true values are for the factors. The best choice one can make is the factor with the highest S_i, whether the model is additive or not and whether the factors are independent or not.

Factors' Fixing (FF) Setting: Can one fix a factor [or a subset of input factors] at any given value over their range of uncertainty without reducing significantly the variance of the output? One can only fix those (sets of) factors whose S_{Ti} is zero.

The extended variance-based methods, including the improved version of Sobol', for both dependent and independent input factors, are implemented in the freely distributed software SIMLAB (Saltelli *et al.*, 2004).

7.2 Results

7.2.1 First analysis

The first analysis was run without imputation, i.e. by censoring all countries with missing data. As a result, only 34 countries could in theory be analysed. We further dropped countries from rank (original TAI) 24, Hong Kong, as this is the first country with missing data, and it was preferred to analyse the set of countries whose rank was not altered the omission of missing records. The uncertainty analysis for the remaining 23 countries is given in Figure 7.1 for the ranks, with countries ordered by their original TAI position, going from Finland, rank=1, to Slovenia, Rank=23. The reader will recall that the choice of ranks, instead of composite indicator values, is dictated by the use of the MCA aggregation system. The width of the 5^{th} – 95^{th} percentile bounds, as well as the fact that the ordering the medians (black hyphen) often is at odd with the ordering of the original TAI (grey hyphen) show that the drastic throwing of all uncertainty sources at the problem, including 3 different aggregation system alternative to each other, results in considerable differences between the new and the original TAI, although one still

sees the difference between the group of leaders and that of laggards. If the uncertainty plugged into the system were a true reflection of the status of knowledge and of the (lack of) consensus among experts on how TAI should be built, we would have to conclude that TAI is not a robust measure of country technology achievement.

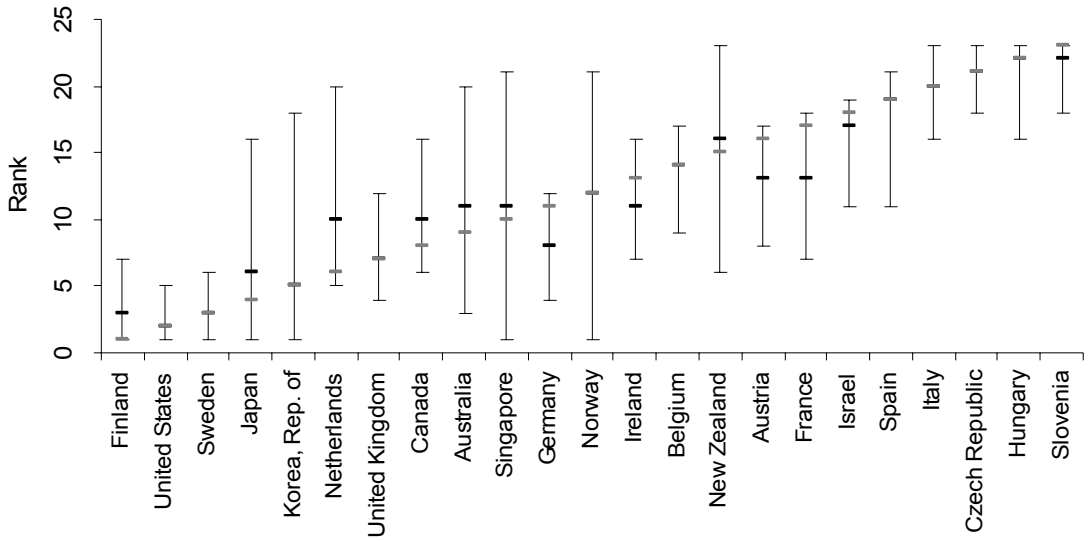


Figure 7.1. *Uncertainty analysis results showing the countries' rank according to the original TAI 2001 (light grey marks), and the median (black mark) and the corresponding 5th and 95th percentiles (bounds) of the distribution of the MC-TAI for 23 countries. Uncertain input factors: normalisation method, inclusion-exclusion of a sub-indicator, aggregation system, weighting scheme, expert selection. Countries are ordered according to the original TAI values.*

Keeping up with this example, we show in Figure 7.2 a sensitivity analysis based on the first order indices calculated using the method of Sobol' (1993) in its improved version due to Saltelli (2002). In fact we present the total variance for each country's rank and how much of it can be decomposed according to the first order conditional variances. We can roughly say that aggregation system, followed by the inclusion-exclusion of sub-indicator and expert selection are the most influential input factors. The countries with the highest total variance in ranks are the middle-performing countries in Figure 7.1, while the leaders and laggards in technology achievement present low total variance. The non-additive, non-linear part of the variances that is not explained by the first order sensitivity indices ranges from 35% for the Netherlands to 73% for United Kingdom, whilst for most countries it exceeds 50%. This underlines the necessity for computing higher order sensitivity indices that capture the interaction effects among the input factors.

Figure 7.3 shows the total effect sensitivity indices for the variances of each country's ranks. The total effect sensitivity indices concentrate in one single term all the interactions involving each input factor and they clearly add up to a number greater than one due to the existing interactions. Again interactions seem to exist among the influential factors already identified.

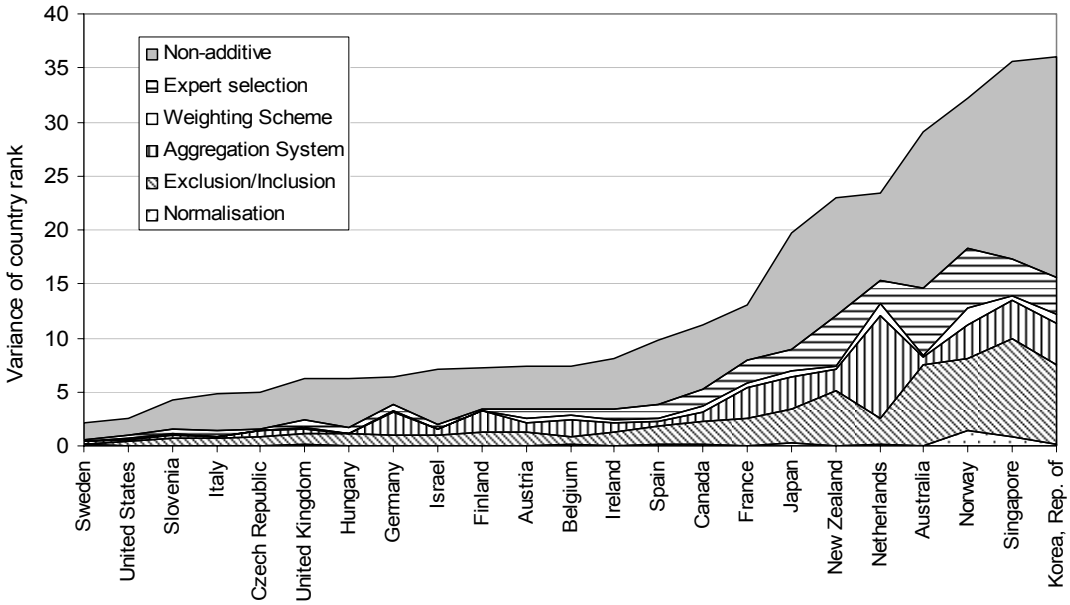


Figure 7.2. *Sensitivity analysis results based on the first order indices. Decomposition of country's variance according to the first order conditional variances. Aggregation system, followed by the inclusion-exclusion of sub-indicator and expert selection are the most influential input factors. The part of the variance that is not explained by the first order indices is noted as non-additive. Countries are ordered in ascending order of total variance.*

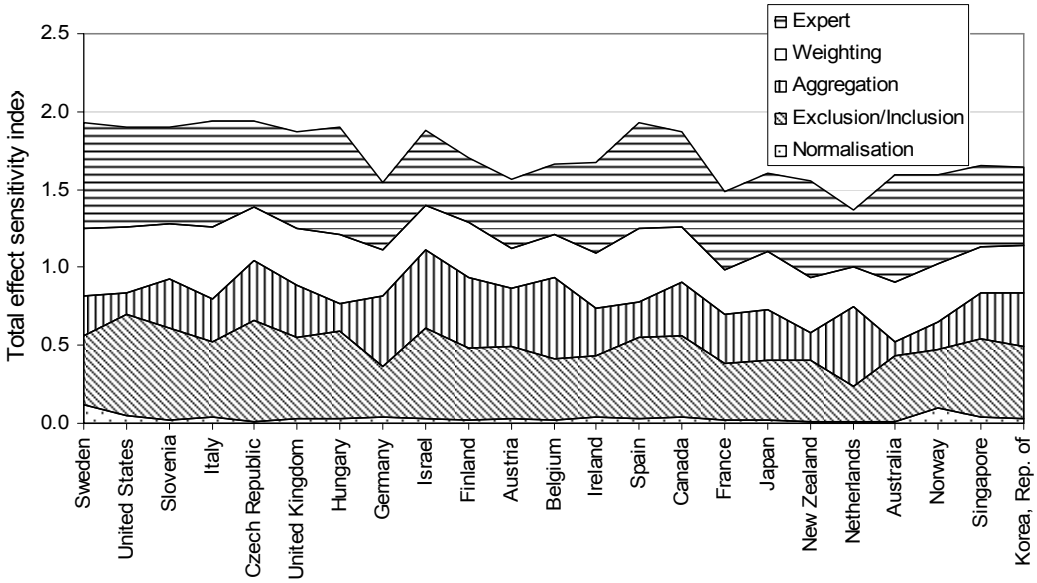


Figure 7.3. *Sensitivity analysis results based on the total effect indices. Aggregation system inclusion-exclusion of sub-indicator and expert selection present most of the interaction effects. Countries are ordered in ascending order of total variance.*

If the TAI model was additive with no interactions between the input factors, the non-additive part of the variance in Figure 7.2 would have been zero (in other words the first order sensitivity indices would have summed to 1) and the sum of the total effect sensitivity indices in Figure 7.3 would have been 1. Yet, the sensitivity indices show the high degree of non linearity and additivity for the TAI model, and of the importance of the interactions. For instance, the high effect of interactions for Netherlands, which also had a large percentile bounds, can be further explored. In Figure 7.4 we see that this country is favoured by combination of "geometric mean system" with "BAL weighing" and unfavoured by combination of "Multi criteria system" with "AHP weighting". This is a clear interaction effect. In depth analysis of the output data reveals that as far as inclusion – exclusion is concerned, it is the exclusion of the sub-indicator "Royalties" leading to worse ranking for the Netherlands under any aggregation system.

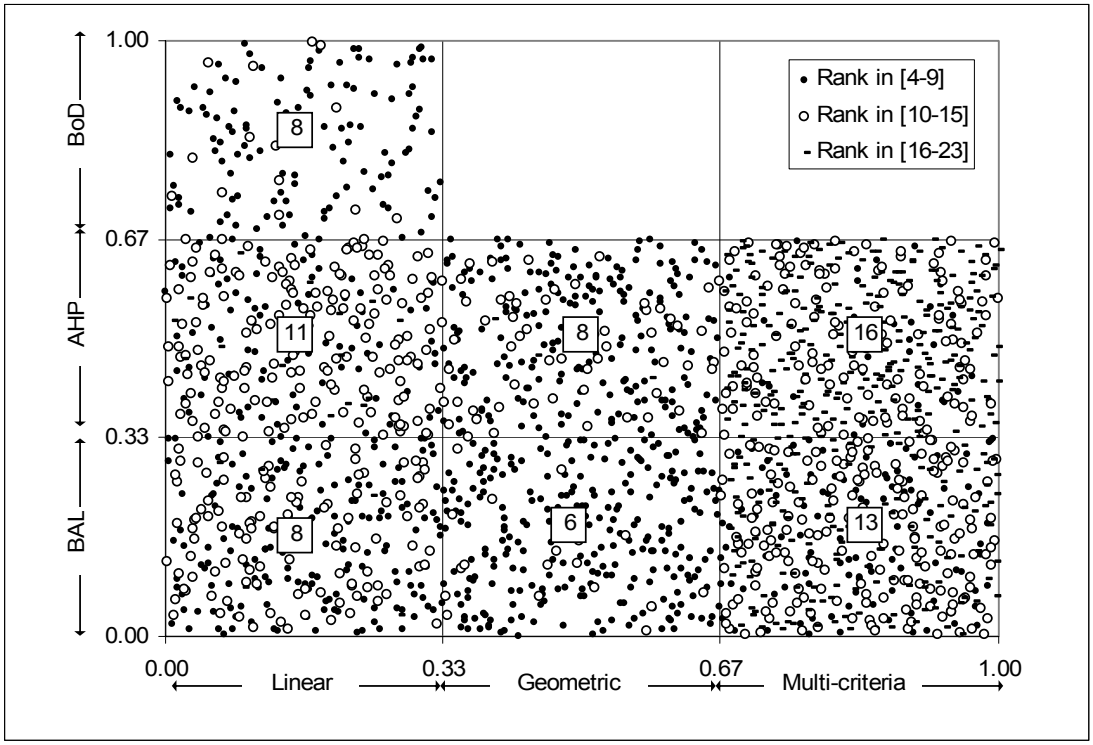


Figure 7.4. *Rank position of the Netherlands for different combinations of aggregation system and weighting scheme. Average rank per case is indicated in the box. The interaction effect between aggregation system and weighting scheme is clear.*

Coming to the output variable average shift in rank (Equation 7.2) with respect to the original TAI rank we see in Figure 7.5 the histogram of the values. The mean value is almost 3 positions, with a standard deviation slightly above 1 position. The input factors affecting this variable the most are aggregation system plus inclusion – exclusion at the first order (Table 7.1), while if the interactions are considered both weighting scheme and expert choice become important (Table 7.1). This effect can be seen in Figure 7.6 where the effect of MCA in spreading the countries ranks can be appreciated. In some cases the average shift in country's rank when using MCA can be as high as 9 places.

Table 7.1. *Sobol' sensitivity measures of first order and total effect for the output: Average shift in countries' rank with respect to the original TAI. Significant values are underlined.*

Input Factors	First order (S_i)	Total effect (S_{Ti})	$S_{Ti} - S_i$
Normalisation	0.000	0.008	0.008
Exclusion/Inclusion of sub-indicator	0.148	<u>0.435</u>	<u>0.286</u>
Aggregation system	<u>0.245</u>	0.425	0.180
Weighting Scheme	0.038	<u>0.327</u>	<u>0.288</u>
Expert selection	0.068	<u>0.402</u>	<u>0.334</u>
Sum	0.499	1.597	

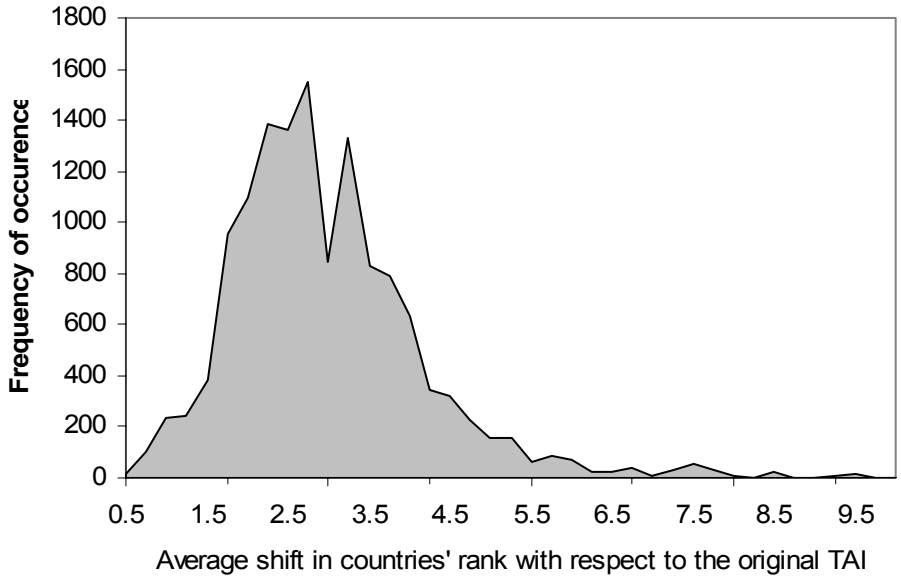


Figure 7.5. *Result of UA for the output variable: average shift in countries' rank with respect to the original TAI. Uncertain input factors: normalisation method, inclusion-exclusion of a sub-indicator, aggregation system, weighting scheme, expert selection.*

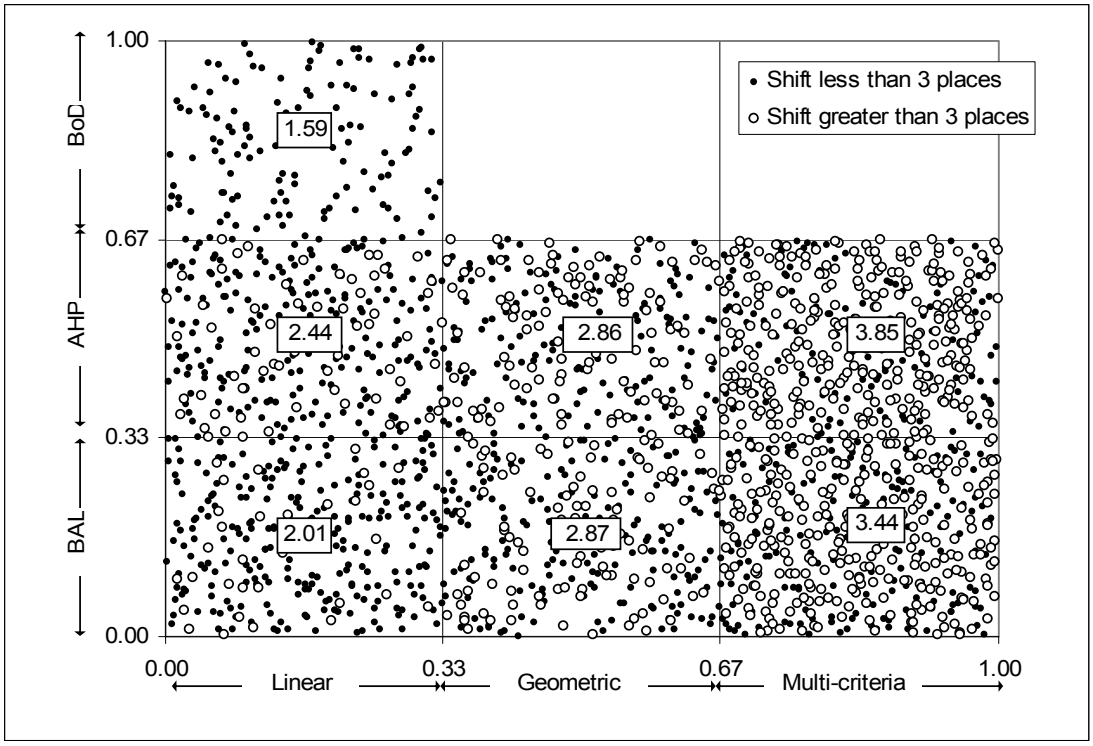


Figure 7.6. *Average shift in countries' rank with respect to the original TAI for different combinations of aggregation system and weighting scheme. Average value per case is indicated in the box.*

7.2.2 Second analysis

As explained above, in the second analysis we make the realistic assumptions that TAI stakeholders have eventually converged to an aggregations system, which we take to be the linear one. In fact, one might argue that the choice of the aggregation system is somehow dictated by the use of the index and by the expectation of its stakeholders. MCA would be adopted if stakeholders were to believe that the system should be non compensatory, e.g. that one unit down on one sub-indicator should not be compensated by a unit up in another. Eventually, this would lead to an on average medium–good performance being worth more for a country than a performance which is very good on some sub-indicators and bad in others. A GME approach would flag the intention to follow progresses of the index over time in a scale-independent fashion (see discussion in section 6).

Based on these considerations, we have based this second analysis on the LIN system, as in the original TAI. The uncertainty analysis plot (Figure 7.7) shows now a much more robust behaviour of the index, with fewer inversion of ranking when median-TAI and original TAI are compared. As far as the sensitivity is concerned, the consideration of uncertainty arising from imputation does not seem to make a significant contribution to the output uncertainties, which are also in this case dominated by weighing, inclusion-exclusion, expert selection. Even when, as in the case of Malaysia, imputation by bivariate approach ends into an unrealistic number of patens being imputed for this country (234 patents granted to residents per million people), its rank's uncertainty is insensitive to imputation. The sensitivity analysis results for the average shift in

ranking output variable (Equation 7.2) is shown in Table 7.2. Interactions are now between expert selection and weighing, and considerably less with interaction with inclusion-exclusion.

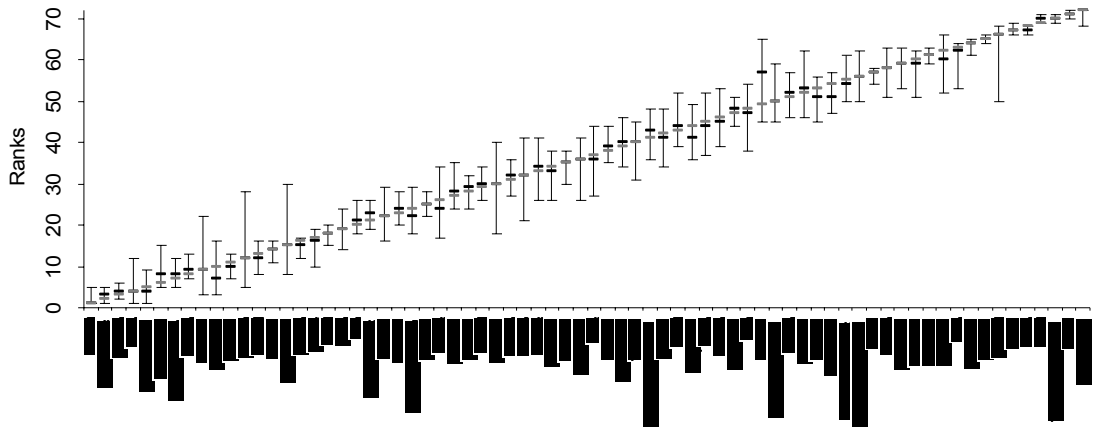


Figure 7.7. *Uncertainty analysis results showing the countries' rank according to the original TAI 2001 (light grey marks), and the median (black mark) and the corresponding 5th and 95th percentiles (bounds) of the distribution of the MC-TAI for 72 countries. Uncertain input factors: imputation, normalisation method, inclusion-exclusion of a sub-indicator, weighting scheme, expert selection. A linear aggregation system is used. Countries are ordered according to the original TAI values*.

Table 7.2. *Sobol' sensitivity measures of first order and total effect for the output: Average shift in countries' rank with respect to the original TAI. Significant values are underlined*.

Input Factors	First order (S_i)	Total effect (S_{Ti})	S_{Ti} - S_i
Imputation	0.001	0.005	0.004
Normalisation	0.000	0.021	0.021
Exclusion/Inclusion of sub-indicator	0.135	0.214	0.078
Weighting Scheme	<u>0.212</u>	<u>0.623</u>	<u>0.410</u>
Expert selection	<u>0.202</u>	<u>0.592</u>	<u>0.390</u>
Sum	0.550	1.453	

7.3 Conclusions

We now go back to our questions on the effect of uncertainties:

(a) Does the use of one strategy versus another in indicator building (the steps i to vii described at the beginning of this section) provide a biased picture of the countries' performance? How this compare to the original TAI index?

The answer to this question is that much depends on the severity of the uncertainties. As shown by our two analyses, if the builders of the index disagree on the aggregation system, there is not much hope that a robust index will emerge, not even by the best provision of uncertainty and sensitivity analysis. If uncertainties exist in the context of a well established theoretical approach,

e.g. the index builder favour a participatory approach within a linear aggregation scheme, then the analysis shows that the countries ranking is fairly robust in spite of the uncertainties.

(b) To what extent do the uncertain input factors (used to generate the alternatives i to vii above) affect the countries' ranks with respect to the original, deterministic TAI?

Both imputation and normalisation do not affect significantly countries ranking when uncertainties of higher order are present. In this present exercises the uncertainties of higher order were expert selection and weighing scheme (second analysis). *A fortiori* normalisation does not affect output when the very aggregation system is uncertain (first analysis). In other words, and generalising these results, when the weights are uncertain, it is unlikely that normalisation and editing will affect sensibly the country ranks.

As discussed above the aggregation system is of paramount importance and it is recommended that indicators builder agree on a common approach. Once the system is fixed, then it is the choice of the aggregation methods and of the experts that – together with indicator inclusion – exclusion, dominates the uncertainty in the country ranks. It is important to mention that even in the second analysis, when the aggregation system is fixed, the composite indicator model is strongly non additive, which reinforces the case for the use of quantitative, Monte Carlo based approaches to robustness analysis.

8. Visualisation

The way composite indicators should be presented is not a trivial issue. Composite indicators must be able to communicate the picture to decision-makers and users quickly and accurately. Visual models of these composites must provide signals, in particular, warning signals that flag for decision-makers those areas requiring policy intervention.

Hereafter we give some interesting ways to display and visualize composite indicators. We accompany each type of visualization by a brief commentary of the pros and cons. We start from the simplest tools and we explore their modifications. We also give reference to the sources that employ these tools.

8.1 Tabular format

This is the simplest format whereby, for each country, the composite indicator and its underlying indicators are presented as a table of values. Usually countries are displayed in decreasing ranking order. An example is the Human Development Index 2004 of the UNDP (see Figure 8.1). This is a comprehensive approach to display results, yet not particularly visually appealing. The approach could be adapted to show targeted information for sets of countries grouped, for example, by location, GDP, etc.

1 Human development index

MONITORING HUMAN DEVELOPMENT: ENLARGING PEOPLE'S CHOICES . . .

HDI rank [a]		Life expectancy at birth (years) 2002	Adult literacy rate (% ages 15 and above) 2002 [b]	Combined gross enrolment ratio for primary, secondary and tertiary schools (%) 2001/02 [c]	GDP per capita (PPP US$) 2002	Life expectancy index	Education index	GDP index	Human development index (HDI) value 2002	GDP per capita (PPP US$) rank minus HDI rank [d]
High human development										
1	Norway	78.9	.. [e]	98 [f]	36,600	0.90	0.99	0.99	0.956	1
2	Sweden	80.0	.. [e]	114 [g,h]	26,050	0.92	0.99	0.93	0.946	19
3	Australia	79.1	.. [e]	113 [g,h]	28,260	0.90	0.99	0.94	0.946	9
4	Canada	79.3	.. [e]	95 [f]	29,480	0.90	0.98	0.95	0.943	5
5	Netherlands	78.3	.. [e]	99 [f]	29,100	0.89	0.99	0.95	0.942	6
6	Belgium	78.7	.. [e]	111 [f,g]	27,570	0.90	0.99	0.94	0.942	7
7	Iceland	79.7	.. [e]	90 [f]	29,750	0.91	0.96	0.95	0.941	1
8	United States	77.0	.. [e]	92 [h]	35,750	0.87	0.97	0.98	0.939	-4
9	Japan	81.5	.. [e]	84 [h]	26,940	0.94	0.94	0.93	0.938	6
10	Ireland	76.9	.. [e]	90 [f]	36,360	0.86	0.96	0.98	0.936	-7
11	Switzerland	79.1	.. [e]	88 [f]	30,010	0.90	0.95	0.95	0.936	-4
12	United Kingdom	78.1	.. [e]	113 [f,g]	26,150	0.88	0.99	0.93	0.936	8
13	Finland	77.9	.. [e]	106 [f,g]	26,190	0.88	0.99	0.93	0.935	6
14	Austria	78.5	.. [e]	91 [f]	29,220	0.89	0.96	0.95	0.934	-4
15	Luxembourg	78.3	.. [e]	75 [f,i]	61,190 [j]	0.89	0.91	1.00	0.933	-14
16	France	78.9	.. [e]	91 [f]	26,920	0.90	0.96	0.93	0.932	0
17	Denmark	76.6	.. [e]	96 [f]	30,940	0.86	0.98	0.96	0.932	-12
18	New Zealand	78.2	.. [e]	101 [g,h]	21,740	0.89	0.99	0.90	0.926	6
19	Germany	78.2	.. [e]	88 [h]	27,100	0.89	0.95	0.94	0.925	-5
20	Spain	79.2	97.7 [e,f,k]	92 [h]	21,460	0.90	0.97	0.90	0.922	5
21	Italy	78.7	98.5 [e,f,k]	82 [f]	26,430	0.89	0.93	0.93	0.920	-3
22	Israel	79.1	95.3	92	19,530	0.90	0.94	0.88	0.908	5
23	Hong Kong, China (SAR)	79.9	93.5 [f,k]	72	26,910	0.91	0.86	0.93	0.903	-6
24	Greece	78.2	97.3 [e,f,k]	86 [f]	18,720	0.89	0.95	0.87	0.902	5
25	Singapore	78.0	92.5 [l]	87 [m]	24,040	0.88	0.91	0.92	0.902	-3

Figure 8.1. *Human Development Index as from the Human Development Report 2004 of the UNDP. The top 25 countries, with high human development, are reported here.*

8.2 Bar charts

The composite indicator is expressed via a bar chart (see Figure 8.2). The countries are on the vertical axis, the values of the composite on the horizontal axis. The top bar indicates the average performance of all countries in the world, and enables the reader to identify how a country is performing with regards to the average.

This figure is used in the publication "Sustainable development indicators in your pocket 2004" of the UK government, a selection of UK indicators of sustainable development (see http://www.sustainable-development.gov.uk/indicators/sdiyp/index.htm).

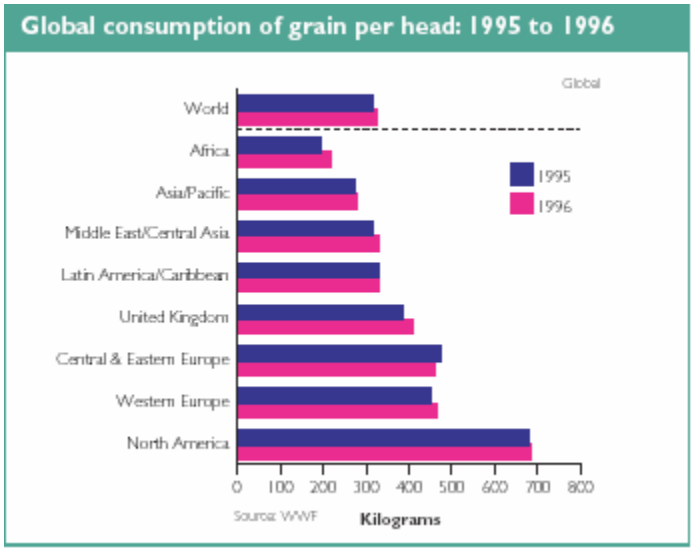


Figure 8.2. *Global consumption of grain per head in two consecutive years.*

The tool is clear, easy to understand. Country comparisons can be made with the average performance. Each underlying sub-indicator can be displayed with a bar chart. The use of colors can make the graph more visually appealing and highlight countries performing well or bad, or showing either growth or slow down, or, finally, to highlight countries having reached an average or mandatory standard. The top bar could alternatively be thought as a target to be reached by countries instead of the current world average. The bar charts show values at two given points in time.

8.3 Line charts

Line charts are used to show performance across time. Performance can be displayed using a) absolute levels; b) absolute growths (in percentage points with respect to the previous year or a number of past years); c) indexed levels and d) indexed growths.

The word 'indexed' means that the values of the indicator are linearly transformed so that their indexed value at a given year is 100. For instance, the indicator called 'Price level index' shows values such that EU15=100 at each year; more expensive countries have values larger than 100, countries cheaper than EU15 have values smaller than 100 (see Figure 8.3).

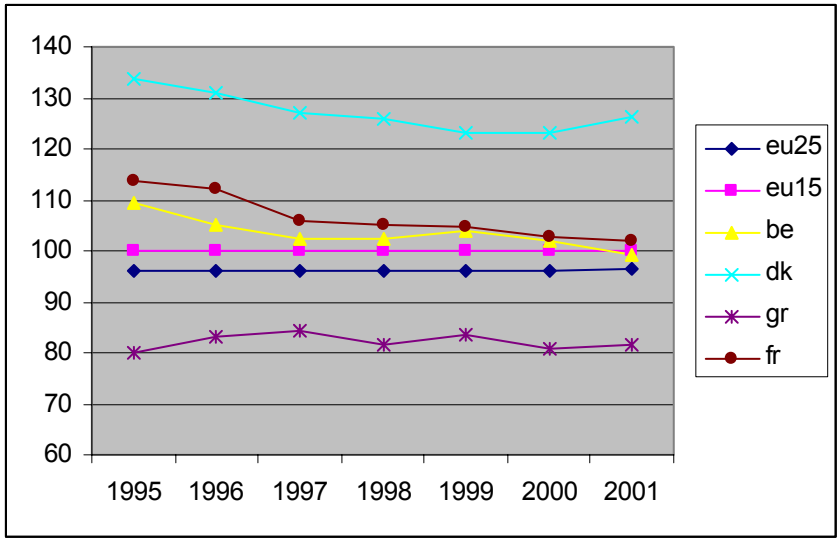


Figure 8.3. *Comparative price levels of final consumption by private households including indirect taxes (EU-15=100). Source: Eurostat. Data retrieved on 4 October, 2004.*

A number of lines are usually superimposed in the same chart to allow comparisons between countries. Another example is given by the Internal Market Index 2004, published on the Internal Market Scoreboard N. 13 (EU-DG MARKT, 2004). Here, groups of countries with similar performance (better, similar or worse than the EU) have been displayed in the same chart. All the countries have been indexed to 100 in the starting year (1994). See an example in Figure 8.4.

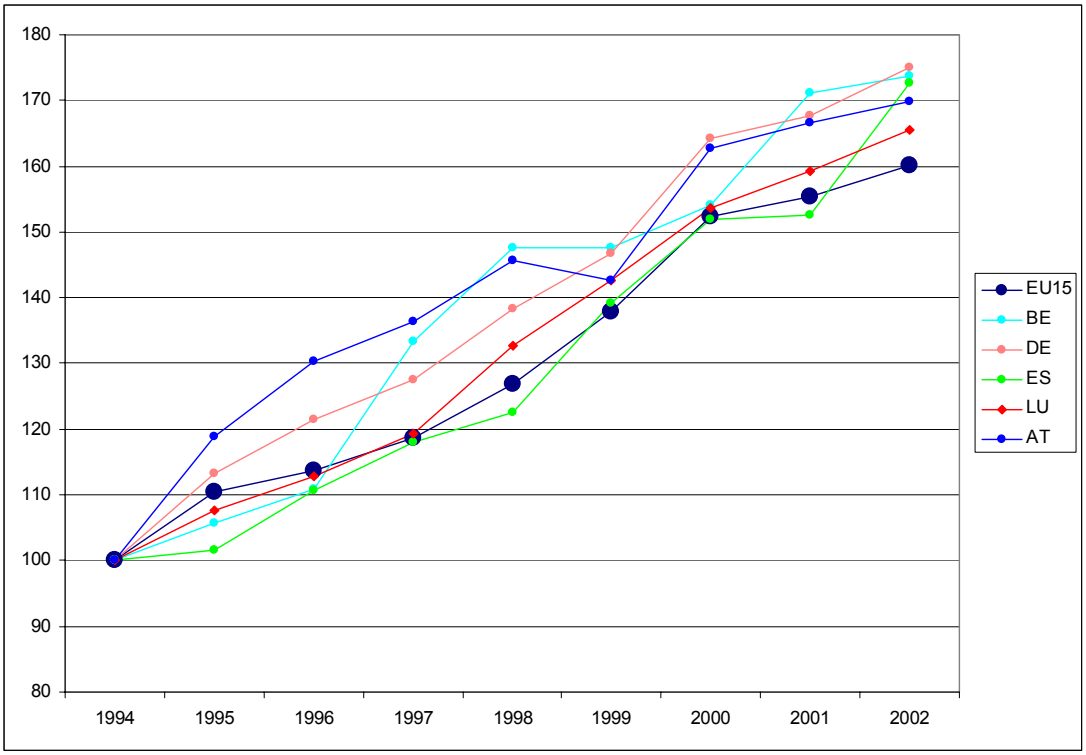


Figure 8.4. *The Internal Market Index for Belgium, Germany, Spain, Luxembourg and Austria improved significantly more than the EU15 average since 1994.*

One can also consider a target for the underlying indicators and add it to the plot. The corresponding target for the composite indicator can be computed and displayed in the plot. See an example in Figure 8.5 taken from Adriaanse, 1993. See also:

http://www.icsu-scope.org/downloadpubs/scope58/box4b.html

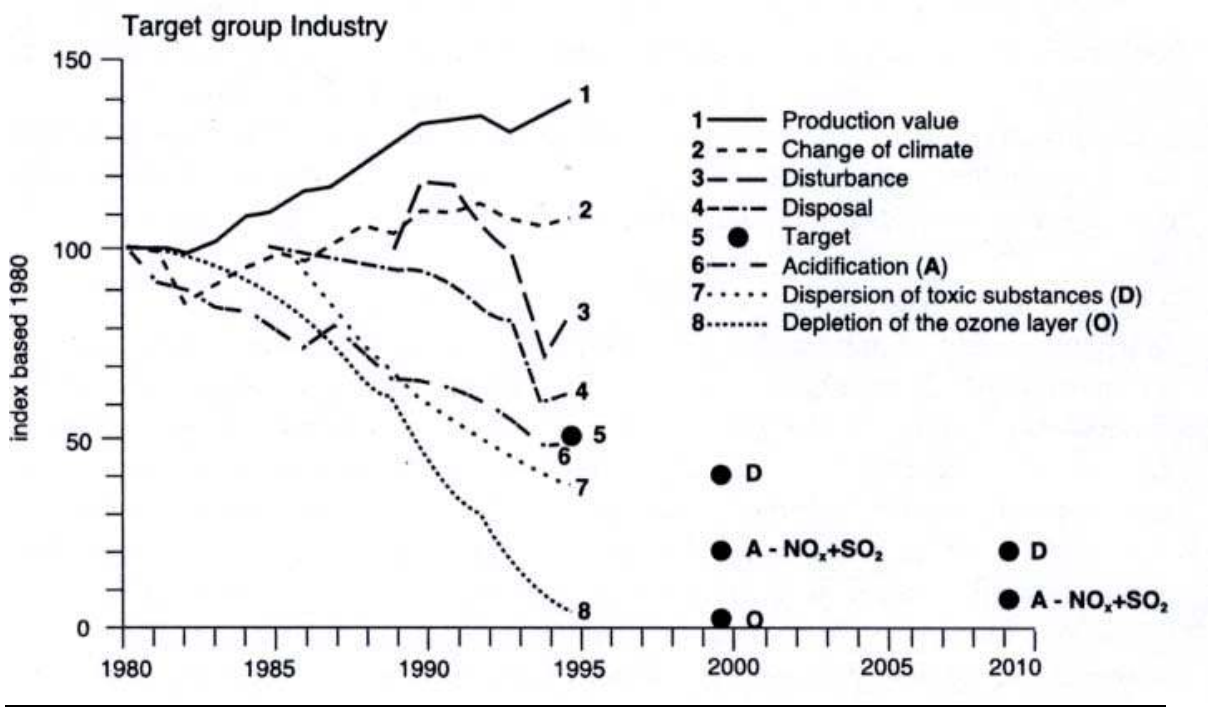


Figure 8.5. *Indicator for the target group industry*

8.4 Traffic lights to monitor progress

For each indicator, where possible, an assessment of progress can be made by comparing the latest data with the position at a number of baselines. Table 8.1 illustrates the approach used by the UK government in sustainable development for three baselines: since 1970, since 1990, and since late 1990s. The 'Traffic light' assessments are used as in Table 8.2.

Table 8.1. *Assessment of each sustainability indicator for three different baselines, as used by UK government.*

Assessment for indicator against objective				
		Change since 1970	**Change since 1990**	**Change since Strategy**[1]
H1 Economic output				
H2 Investment				
H3 Employment				
H15 Waste	All arisings and management			
	Household waste			

Table 8.2. *Traffic-light assessments used by UK government in sustainable development*

Key	
	Significant change, in direction of meeting objective
	No significant change
	Significant change, in direction away from meeting objective
	Insufficient or no comparable data

8.5 Rankings

A quick and easy way to display country performance is to use rankings. It consists in a simple tabular representation such as that supplied by the *Growth Competitiveness Index*, in the Global Competitiveness Report 2003-2004 published by the World Economic Forum (see Figure 8.6). The table shows the rankings of countries for two consecutive years. Thus, it can be used to track changes of country performance over time. The limitation of ranks is that one loses the information on the difference between countries performances.

GROWTH COMPETITIVENESS INDEX RANKINGS

Country	Growth Competitiveness ranking 2003	Growth Competitiveness ranking 2003 among GCR 2002 countries	Growth Competitiveness ranking 2002*
Finland	1	1	1
United States	2	2	2
Sweden	3	3	3
Denmark	4	4	4
Taiwan	5	5	6
Singapore	6	6	7
Switzerland	7	7	5
Iceland	8	8	12
Norway	9	9	8
Australia	10	10	10
Japan	11	11	16
Netherlands	12	12	13
Germany	13	13	14
New Zealand	14	14	15
United Kingdom	15	15	11
Canada	16	16	9
Austria	17	17	18
Korea	18	18	25
Malta	19	—	—
Israel	20	19	17
Luxembourg	21	—	—
Estonia	22	20	27

Figure 8.6. *Growth competitiveness index rankings from the Global competitiveness Report 2003-2004.*

8.6 Scores and rankings

In several cases one provides both levels and country rankings, for both the sub-indicators and the composite one. The British Office of National Statistics has produced indices of economic deprivation in six domains (income, employment, health deprivation and disability; education; skills and training; housing; and access to services) for the all the districts in 2000. The composite

is the average of scores out of a 100 for each sub-indicator (see Table 8.3). The rank is the average of ranks for each sub-indicator; ranks go from 1 to approximately 8,000 (the total number of districts).

Table 8.3. *Index of multiple deprivation by district in England, Office of National Statistics*

Variable	**Index of multiple deprivation**					
Units	**Score**					
Area			**Income domain**		**Employment domain**	
Ascot	5.20	**Rank**	**Score**	**Rank**	**Score**	**Rank**
Binfield	5.13					
Bullbrook	18.72	7,991	7.38	7,640	2.26	8,330
Central Sandhurst	6.55	8,014	5.36	8,205	1.80	8,388
College Town	4.18	3,811	19.23	3,198	8.46	4,087
Cranbourne	12.70	7,614	11.62	5,746	3.36	7,923
Crowthorne	10.32	8,188	4.64	8,300	2.53	8,284
Garth	15.14	5,460	12.29	5,443	4.56	7,100
Great Hollands North	12.55	6,256	5.04	8,257	8.32	4,177
Great Hollands South	12.28	4,690	15.68	4,200	7.54	4,669
Hanworth	10.75	5,517	17.81	3,574	5.69	6,156

Scores and moving average

Sometimes we want to monitor not only the performance at a given point in time but also the trend over the last period. Very often this is done via the calculation of percentage growth, yet moving average can be a useful tool.

An example is given by First Great Western Link railways, which use this tool to inform the public about the punctuality of the Thames trains service. One can read the most recent figure on punctuality and the corresponding moving average over the last 52 weeks. If the moving annual average over the last 12 months for punctuality is less than an appropriate threshold, a discount of up to 5% will be given on qualifying season ticket renewals!

Four-quadrant model for sustainability

Arup (a professional consultancy group) developed a tool to demonstrate the sustainability of a project, process or product to be used either as a management information tool or as part of a design process. The Sustainable Project Appraisal Routine (SPeAR®) is based on a four-quadrant model that structures the issues of sustainability into a robust framework, from which an appraisal of performance can be undertaken (see Figure 8.7). The outcome of the SPeAR® assessment reflects the utilisation of an unweighted indicator set. SPeAR® contains a set of core sectors and indicators that have been derived from the literature on sustainability. The appraisal is based on the performance of each indicator against a scale of best and worst cases. Each indicator scenario

is aggregated into the relevant sector and the average performance of each sector is then transferred onto the SPeAR® diagram. The transparent methodology behind the SPeAR® diagram ensures that all scoring decisions are fully audit traceable. The only limitation is that the diagram gives snapshot of performance at a particular time.

Figure 8.7. *The four-quadrant model of the Sustainable Project Appraisal Routine (SPeAR®).*

8.7 Dashboards

The Dashboard of Sustainability (see http://esl.jrc.it/envind/) is a free, non-commercial software which allows to present complex relationships between economic, social and environmental issues in a highly communicative format aimed at decision-makers and citizens interested in Sustainable Development. It is also particularly recommended to students, university lecturers, researchers and indicator experts.

The Dashboard includes maps of all continents and can be developed using one's own dataset. A vast collection of dashboards already exist. To make some examples, on the internet site one can find the "ecological footprint", a pure environmental composite, the "environment sustainability index", presented by the World Economic Forum annual meetings, the "European Environmental Agency's EEA Environmental Signals". The "From Rio to Johannesburg" and the "Millennium Development Goals" versions are recommended for introductory courses on Sustainable Development.

The Dashboard can help answering some typical questions as:

1. What is the situation of my country compared to others (see Figure 8.8)?
2. What are specific strengths and weaknesses of my continent/my country (Figure 8.9)?
3. How are certain indicators linked to each other (Figure 8.10)?

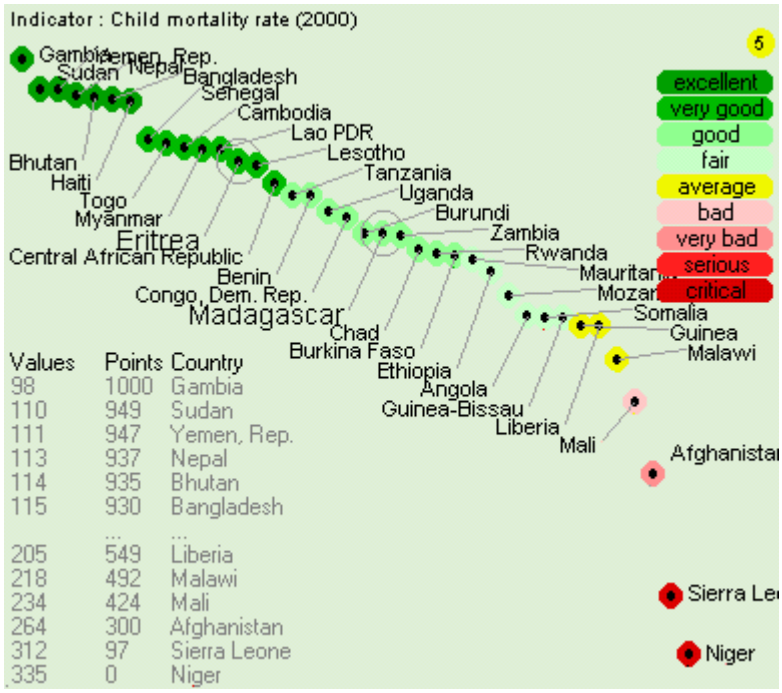


Figure 8.8. *What is the situation of my country compared to others? Source: Dashboard of Sustainability*

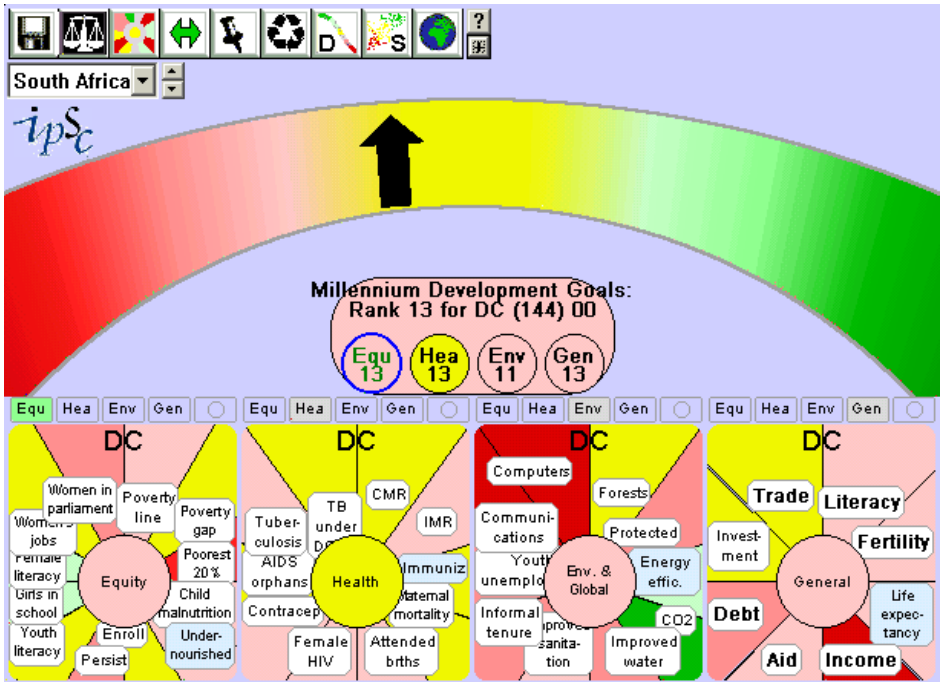


Figure 8.9 *What are specific strengths and weaknesses of my continent/my country? Source: Dashboard of Sustainability*

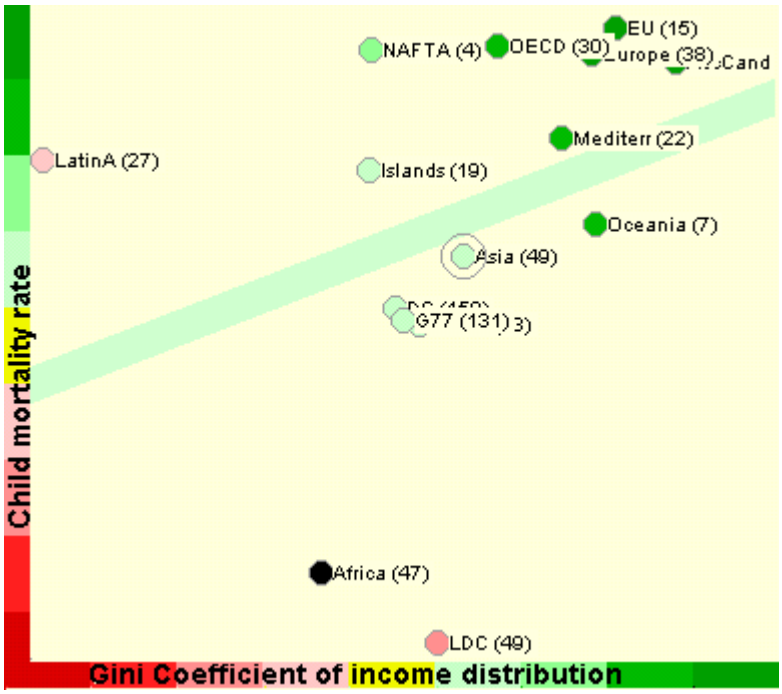


Figure 8.10. *How are certain indicators linked to each other? Source: Dashboard of Sustainability*

8.8 Nation Master

The following internet site is not strictly for composite indicators. However its graphical features can be helpful for presentational purposes.

www.nationmaster.com is a massive central data source on the internet with a handy way to graphically compare nations. Nation Master is a vast compilation of data from such sources as the CIA World Factbook, United Nations, World Health Organization, World Bank, World Resources Institute, UNESCO, UNICEF and OECD.

It is possible to generate maps and graphs on all kinds of statistics with ease.

On October 2004, it includes 4,350 stats, and new features and new statistics are constantly added. This internet site is considered the web's one-stop resource for country statistics on anything and everything.

Correlation reports and scatterplots can be used to find relationships between variables. Integrated into these is a full encyclopedia with over 200,000 articles. See Figure 8.11 for a snapshot.

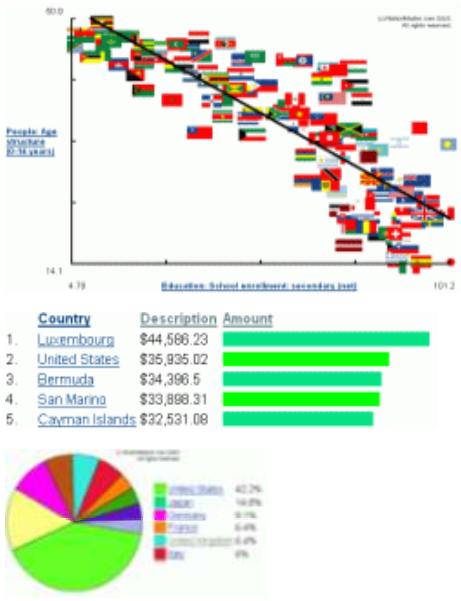


Figure 8.11. *A snapshot from the Nation Master*

Levels vs. growths

When a composite indicator is available for a set of countries for at least two different time points, one is commonly interested not only in the levels at a given time point, yet also in the growths between the available years.

An example is given in the 2003 edition of the European Innovation Scoreboard developed by the European Commission at the request of the Lisbon Council in 2000. The scoreboard includes the

Summary Innovation Index to track relative performance of Member Countries in Innovation. Here overall country trends are reported on the X-axis and levels are given on the Y-axis (see Figure 8.12). A horizontal axis gives the EU average value and a vertical axis gives the EU trend. The two axes divide the area into four quadrants. Countries in the upper quadrant are "moving ahead", because both their value and their trend are above the EU average. Countries in the bottom left quadrant are "falling further behind" because they are below the EU average for both variables. The underlying indicators of innovation are also reported in a separate graph for each country separately. See for example the case of Italy in Figure 8.13, where levels relative to the EU15 countries are reported (with red, yellow and green horizontal bars) as well as the trend relative to EU 15.

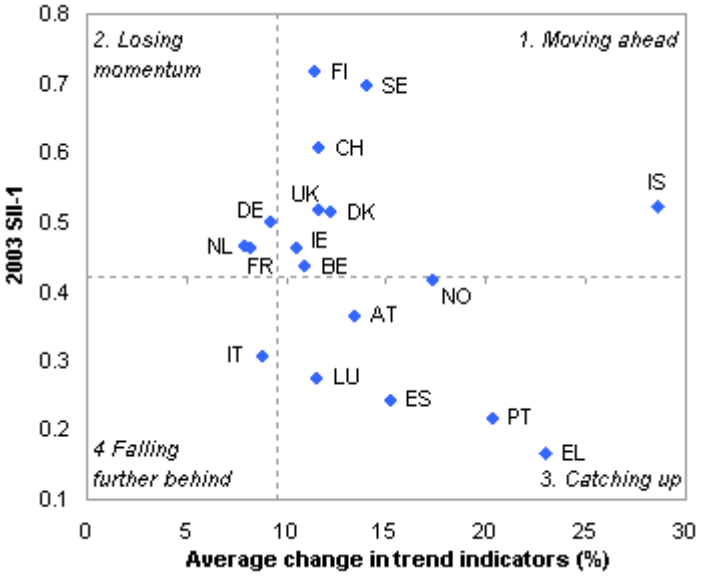


Figure 8.12. *Overall country trend by Summary Innovation Index*

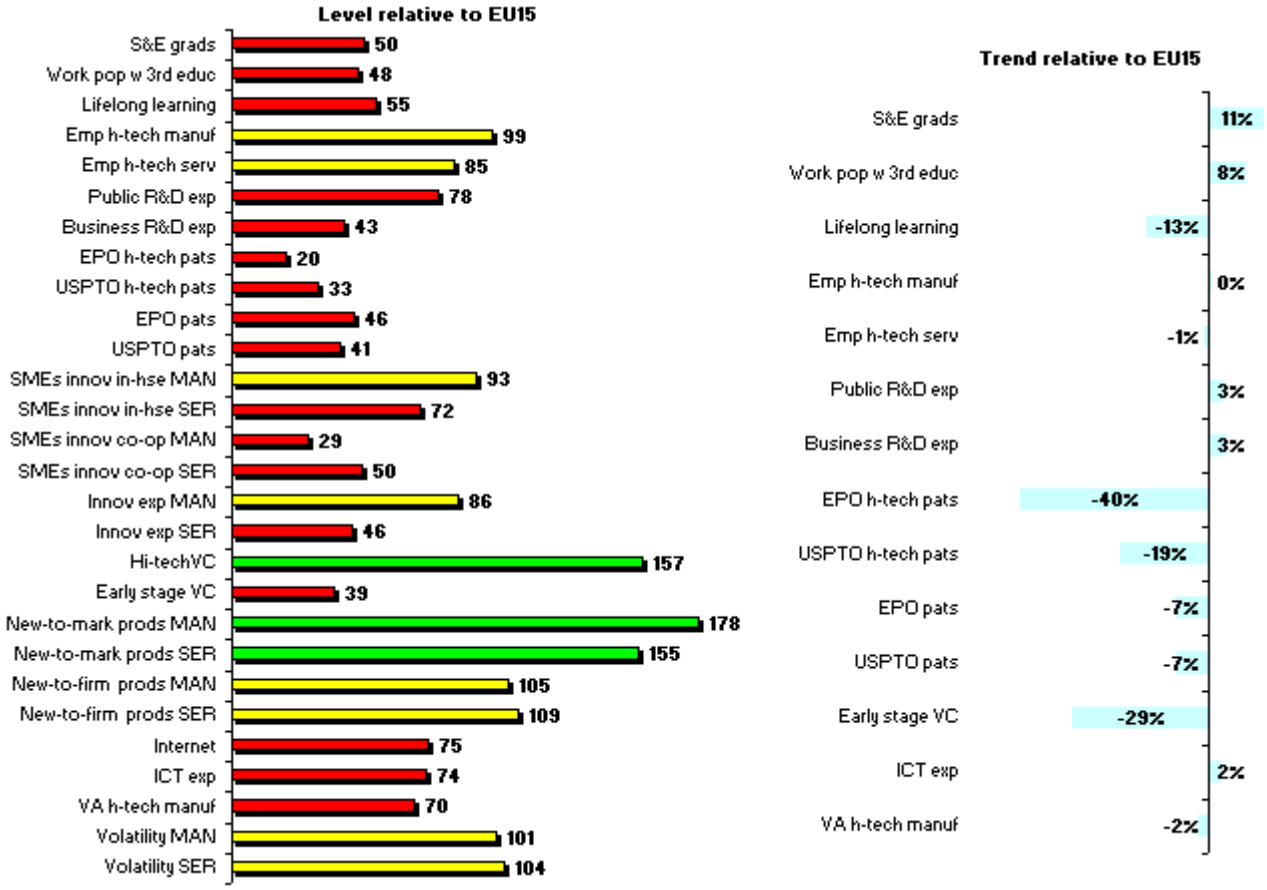
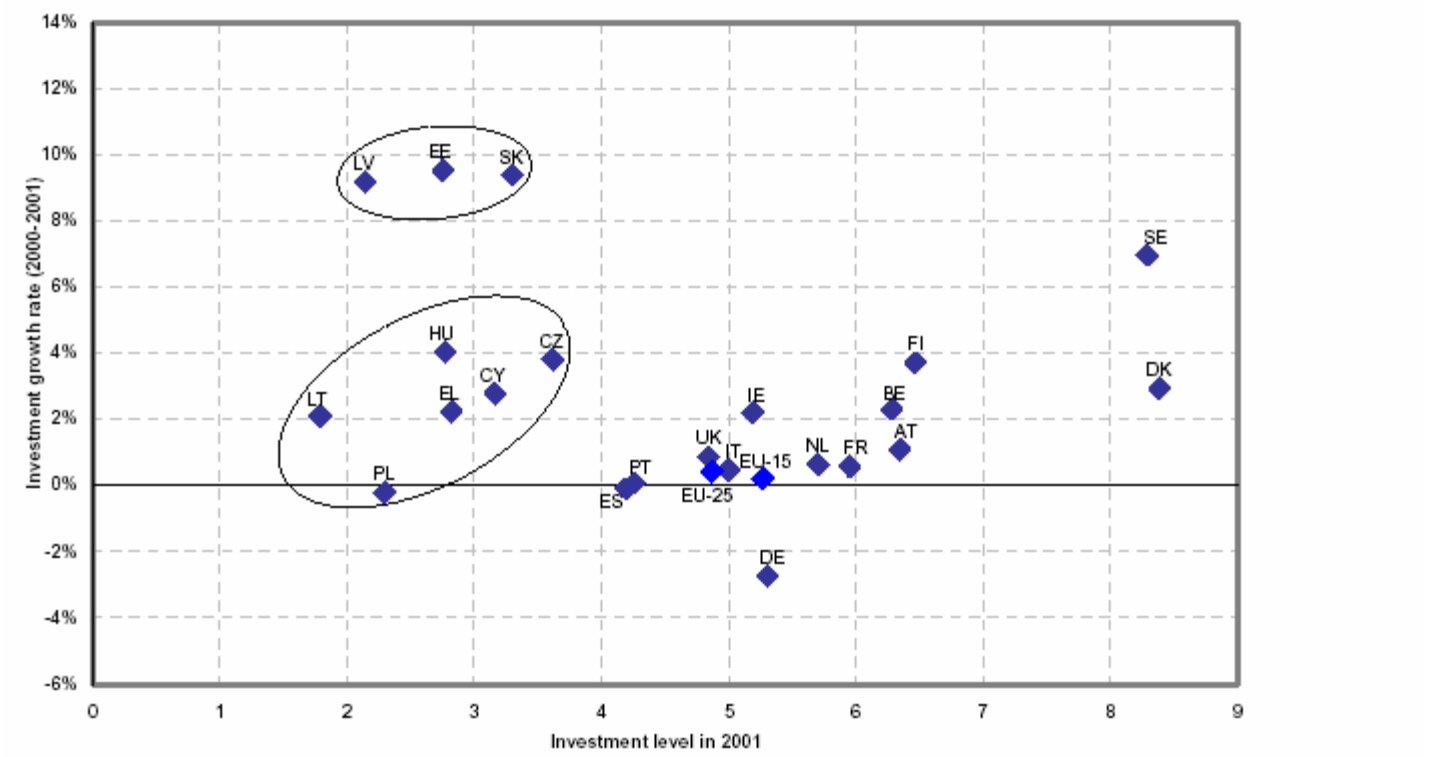


Figure 8.1 *Innovation indicators: performance of Italy relative to EU15 in 2003.*

Another example of this presentational tool is given by the composite indicators of investment and performance in the knowledge-based economy, also developed by the European Commission in the framework of the Lisbon agenda. In the publication *Key Figures 2003/2004* of the Directorate General RTD one can find pictures like that given in Figure 8.14, where levels are given along the X-axis, and short term trends on the Y-axis.



Source: DG Research/JRC Key Figures 2003-2004
Note: Only 5 sub-indicators were included: R&D expenditure (GERD per capita), PhDs (number of new S&T PhDs per capita), Researchers (number of researchers per capita), gross fixed capital formation (GFCF (excluding building) per capita), and e-government. The other two sub-indicators (educational spending and life-long-learning) are not available for all countries. L, MT, SL are not included (no data for most of indicators).

Figure 8.14. *Composite indicator of investment in the knowledge-based economy for comparison between the EU-15 and the former Acceding Countries.*

8.9 Comparing indicators using clusters of countries

In the United Nations Industrial Development Organization (UNIDO) publication *Industrial Development Report 2002/2003: Competing through Innovation and Learning*, (see http://www.unido.org/userfiles/hartmany/12IDR_full_report.pdf), at page 50, the technological evolution of industry in industrialized and transition economies in years 1985 and 1998 is based on clusters of countries with similar performance (see Figure 8.15). This format can be used to plot levels vs. growths for a given composite indicator.

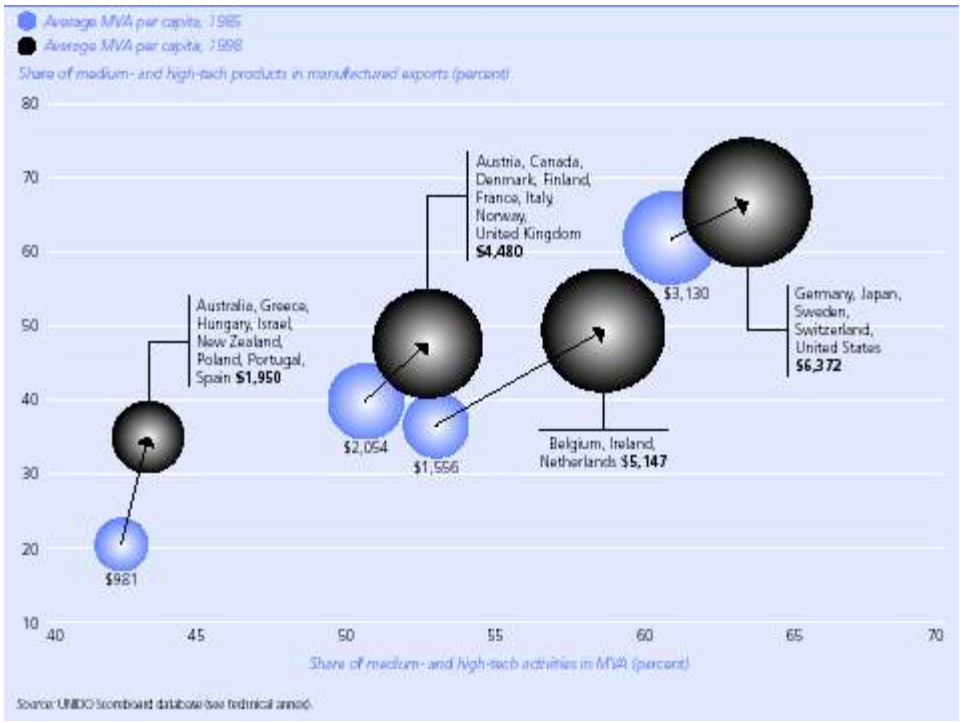


Figure 8.15. *Technological evolution in industry both in industrialized and transition economies in years 1985 (blue cluster) and 1998 (black cluster). Source: UNIDO.*

Graphical representation of composite indicators should provide a clear and identifiable message, but without obscuring the individual data points on which they depend. Booz Allen Hamilton consulting developed a technique of graphical 'profile' indicators to achieve this. The 2003 International Benchmarking Study (IBS) includes a newly devised 'Sophistication Index' designed to provide a deeper insight into the true level of sophistication of a nation's businesses' use of ICT than simple measures of connectivity or adoption.

The chart given in Figure 8.16 lays out all elements of the sophistication index, arranged vertically down the left hand axis. The horizontal scale represents the index score achieved by the UK for each component indicator, normalized between 0 and 1. To score a perfect 1.0, a nation must emphatically lead across all the indicators. For this reason the best performer in the group is generally less than 1. The segmented line represents the composite outcome for the UK across the set of indicators. The scores of the other nations are reported without labels. Only the best performer in each single indicator is given. The approach permits the focus to remain on sharing successful policies.

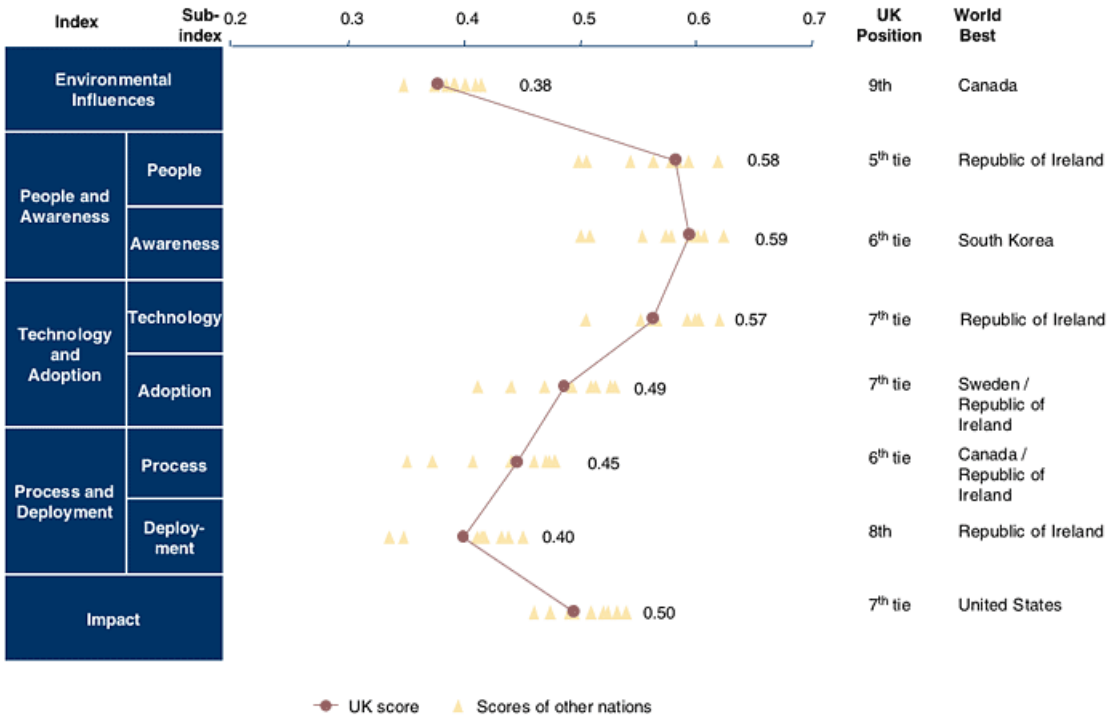


Figure 8.16. *Sophistication Index' proposed by Booz Allen Hamilton to measure a nation's businesses' use of ICT.*

9. Conclusions

Our society is changing so fast that we need to know as soon as possible when things go wrong. Without rapid alert signals, appropriate corrective action is impossible. This is where composite indicators could be used as yardstick. This document provides the readers with a structured way of thinking for the design and construction of composite indicators.

Here is an outline of the issues we touched upon in this report:

- **Theoretical framework** - *What is badly defined is likely to be badly measured.*

The first step in the construction of a composite indicator is the definition and specification of what should be measured, which implies the recognition of the multidimensional nature of the phenomenon to be measured and the effort of specifying the single aspects and their interrelation.

- **Data selection** – *The quality of composite indicators depends also on the quality of the underlying indicators.*

Variables that express different aspects of what is being measured should be selected on the basis of their analytical soundness, measurability, relevance to the phenomenon being measured, and relationship to each other.

- **Multivariate analysis** – *Multivariate statistics is a powerful tool for investigating the inherent structure in the indicators' set.*

This type of analysis is of exploratory nature and is helpful in assessing the suitability of the dataset and providing an understanding of the implications of the methodological choices (e.g. weighting, aggregation) during the construction phase of the composite indicator.

- **Imputation of missing data**– *The idea of imputation is both seductive and dangerous.*

Several imputation methods are available, from the simple regression- to more complicate multiple-imputation. The advantages of imputation include the minimisation of bias and the use of 'expensive to collect' data that would otherwise be discarded. The main disadvantage of imputation is that the results are affected by the imputation algorithm used.

- **Normalisation** – *Avoid adding up apples and pears.*

Normalization serves the purpose of bringing the indicators into the same unit. There are a number of normalization methods available, such as ranking, standardization, re-scaling, distance to reference country, categorical scales, cyclical indicators, balance of opinions. The selection of a suitable normalization method to apply to the problem at hand is not trivial and deserves special care.

- **Weighting and aggregation** – *Relative importance of the indicators and compensability issues*.

Central to the construction of a composite index is the need to combine in a meaningful way the different dimensions, which implies a decision on the weighting model and the aggregation procedure. Different weighting and aggregation rules are possible. Each technique implies different assumptions and has specific consequences.

- **Robustness and sensitivity** – *The iterative use of uncertainty and sensitivity analysis during the development of a composite indicator can contribute to its well-structuring*.

The construction of composite indicators involves stages where judgement has to be made affecting the message brought by the composite indicator in a way that deserves analysis and corroboration. A plurality of methods (all with their implications) should be initially considered, because no model (construction path of the composite indicator) is a priori better than another and because each model serves different interests. Uncertainty and sensitivity analysis are the suggested tools for coping with uncertainty and ambiguity in a more transparent and defensible fashion.

- **Visualisation** – *If arguments are not put into figures, the voice of science will never be heard by practical men.*

Composite indicators must be able to communicate the picture to decision-makers and users quickly and accurately. Visual models of these composite indicators must be able to provide signals, in particular, warning signals that flag for decision-makers those areas requiring policy intervention. The literature presents various ways for presenting the composite indicator results, ranging from simple forms, such as tables, bar or line charts, to more sophisticated figures, such as the four-quadrant model (for sustainability), the Dashboard, etc.

The discussions in this document show how Composite Indicators 'naturally' emerge in a context where country performance is being benchmarked, we discuss some salient aspect of the

composite indicators controversy, pitting "Aggregators" and "Non-Aggregators against one another. For example:

'Composite indicators are confusing entities whereby apples and pears are added up in the absence of a formal model or justification.'

against

'Composite indicators are a way of distilling reality into a manageable form.'

Whether or not one likes or accepts composite indicators for the purpose of comparing countries performance, one might find itself exposed to a composite indicator even when unwilling. And yet the composite indicators controversy is there to stay. But the bottle-neck conclusion is that composite indicators should never be seen as a goal per se, regardless of their quality or underlying variables. They should be seen, instead, as a starting point for initiating discussion and attracting public interest and concern.

REFERENCES and BIBLIOGRAPHY

1. Adriaanse A., (1993) Environmental policy performance. A study on the development of indicators for environmental policy in the Netherlands. SDV Publishers, The Hague.
2. Anderberg, M.R. (1973), Cluster Analysis for Applications, New York: Academic Press, Inc.
3. Arrow K.J. (1963) - Social choice and individual values, 2d edition, Wiley, New York.
4. Arrow K.J., and Raynaud H. (1986) - Social choice and multicriterion decision making, M.I.T. Press, Cambridge.
5. Arundel A. and Bordoy C. (2002) Methodological evaluation of DG Research's composite indicators for the knowledge based economy. Document presented by DG RTD at the Inter-service consultation meeting on Structural Indicators on July 11th 2002.
6. Binder, D.A. (1978), "Bayesian Cluster Analysis," Biometrika, 65, 31 -38.
7. Boscarino J.A., Figley C.R., and Adams R.E., (2004) Compassion Fatigue following the September 11 Terrorist Attacks: A Study of Secondary Trauma among New York City Social Workers, International Journal of Emergency Mental Health, Vol. 6, No. 2 • 2004, 1-10.
8. Box, G., Hunter, W. and Hunter, J. (1978) Statistics for experimenters, New York: John Wiley and Sons.
9. Box G.E.P., (1979), Robustness is the strategy of scientific model building. In R.L. Launer and G.N. Wilkinson (Eds.), Robustness in Statistics, Academic Press New York, pp. 201-236.
10. Bryant F.B., and Yarnold P.R., (1995). Principal components analysis and exploratory and confirmatory factor analysis. In Grimm and Yarnold, Reading and understanding multivariate analysis. American Psychological Association Books.
11. Chan, K., Tarantola, S., Saltelli, A. and Sobol', I. M. (2000) Variance based methods. In Sensitivity Analysis (eds A. Saltelli, K. Chan, M. Scott) pp. 167-197. New York: John Wiley & Sons.
12. Charnes A., Cooper W.W., Lewin A.Y., and Seiford L.M., (1995), Data Envelopment Analysis: Theory, Methodology and Applications. Boston:Kluwer.
13. Cherchye L. (2001), Using data envelopment analysis to assess macroeconomic policy performance, Applied Economics, 33, 407-416.
14. Cherchye L., and Kuosmanen T., (2002), Benchmarking sustainable development: a synthetic meta-index approach, EconWPA Working Papers.
15. Cherchye, L., Moesen, W. and Van Puyenbroeck, T., (2004), "Legitimately Diverse, yet Comparable: on Synthesizing Social Inclusion Performance in the EU", Journal of Common Market Studies 42, 919-955.
16. Commission of the European Communities (1984) The regions of Europe: Second periodic report on the social and economic situation of the regions of the Community, together with a statement of the regional policy committee, OPOCE, Luxembourg.
17. Cortina, J.M. (1993) What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 1, 98-104.
18. Cox, D., Fitzpatrick, R., Fletcher, A., Gore, S., Spiegelhalter, D. and Jones, D. (1992) Quality-of-life assessment: can we keep it simple? J.R. Statist. Soc. **155 (3)**, 353-393.
19. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika. 16, 297-334.
20. Davis, J., (1986); Statistics and Data Analysis in Geology , (John Wiley & Sons, Toronto, 646p.).
21. Pan American Health Organization (1996) Annual report of the Director. Healthy People, Healthy Spaces 1996, Official Document No. 283, Washington, D.C. 20037, U.S.A. http://165.158.1.110/english/sha/ops96arx.htm

22. Debreu G. (1960) - Topological methods in cardinal utility theory, in Arrow K.J., Karlin S. and Suppes P. (eds.) Mathematical methods in social sciences, Stanford University Press, Stanford.
23. Dempster A.P. and Rubin D.B. (1983), Introduction pp.3-10, in Incomplete Data in Sample Surveys (vol. 2): Theory and Bibliography (W.G. Madow, I. Olkin and D.B. Rubin eds.) New York: Academic Press.
24. Dietz F.J.,and van der Straaten J. (1992) - Rethinking environmental economics: missing links between economic theory and environmental policy, Journal of Economic Issues, Vol. XXVI No. 1, pp. 27-51.
25. Dunteman, G.,H. (1989). Principal components analysis. Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 69.
26. European Commission - DG ENTR (2001) European Innovation Scoreboard, Brussels.
27. Ebert U. and Welsch H., (2004), Meaningful environmental indices: a social choice approach, Journal of Environmental Economics and Management, vol. 47, pp. 270-283.
28. Emam K., Goldenson D., McCurley J., and Herbsleb J., (1998) Success or failure? Modeling the likelihood of software process improvement. International software engineering research network, Technical Report ISERN-98-15.
29. Environmental Protection Agency (EPA), Council for Regulatory Environmental Modeling (CREM), Draft Guidance on the Development, Evaluation, and Application of Regulatory Environmental Models", http://www.epa.gov/osp/crem/library/CREM%20Guidance%20Draft%2012_03.pdf.
30. Euroabstracts (2003) *Mainstremaing Innovation*. Published by the European Commission, Innovation Directorate, Vol. 41-1, February 2003.
31. European Commission (2000) Business Climate Indicator, DG ECFIN, European Commission, Brussels.
32. European Commission, (2001a), Summary Innovation Index, DG Enterprise, European Commission, Brussels.
33. European Commission, (2001b), Internal Market Scoreboard, DG MARKT, European Commission, Brussels.
34. European Commission (2004a), Economic Sentiment Indicator, DG ECFIN, Brussels, http://europa.eu.int/comm/economy_finance/index_en.htm
35. European Commission (2004 b), composite Indicator on e-business readiness, DG JRC, European Commission, Brussels.
36. Everitt, B.S. (1979), "Unresolved Problems in Cluster Analysis," Biometrics, 35, 169 -181.
37. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4: 272-299.
38. Fagerberg J. (2001) Europe at the crossroads: The challenge from innovation-based growth in the Globalising Learning Economy, B. Lundvall and D. Archibugi eds., Oxford Press.
39. Feldt, L.S., Woodruffe, D.J., and Salih, F.A. (1987) Statistical Inference for Coefficient Alpha. Applied Psychological Measurement, 11,1, 93-103.
40. Forman E.H. (1983), The analytic hierarchy process as a decision support system, Proceedings of the IEEE Computer society.
41. Freudenberg, M. (2003) Composite indicators of country performance: a critical assessment. Report DSTI/IND(2003)5, OECD, Paris.
42. Funtowicz S.O., Munda G., Paruccini M. (1990) - The aggregation of environmental data using multicriteria methods, Environmetrics, Vol. 1(4), pp. 353-36.
43. Funtowicz S.O., Ravetz J.R. (1990) - Uncertainty and quality in science for policy, Kluwer Academic Publishers, Dordrecht.
44. Giampietro M., Mayumi K., Munda G., (2004), Integrated Assessment and Energy Analysis: Qualitative Assurance in Multi-Criteria Analysis of Sustainability, forthcoming in Energy.

45. Girardin P., Bockstaller C., and Van der Werf H., (2000), Assessment of potential impacts of agricultural practives on the environment: the AGRO*ECO method. Environmental impact Assessment Review, vol.20, pp. 227-239.
46. Gorsuch, R. L. (1983). Factor Analysis. Hillsdale, NJ: Lawrence Erlbaum. Orig. ed. 1974.
47. Gough C., Castells, N., and Funtowicz S., (1998), Integrated Assessment: an emerging methodology for complex issues, Environmental Modeling and Assessment, n.3, 19-29.
48. Green, S.B., Lissitz, R.W., and Mulaik, S.A.(1977) Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 37, 827-838.
49. Green P.E., and Srinivasan V., (1978), Conjoint analysis in consumer research: issues and outlook. Journal of Consumer research 5, 103-123.
50. Grubb D., and Wells W., (1993), Employment regulation and patterns of work in EC countries, OECD Economic Studies, n. 21 Winter, 7-58, Paris.
51. Hair J.F., Anderson R.E., Tatham R.L., and Black W.C., (1995), Multivariate data Analysis with readings, fourth ed. Prentice Hall, Englewood Cliffs, NJ.
52. Hartigan, J.A. (1975), Clustering Algorithms, New York: John Wiley & Sons, Inc.
53. Harvey A., (1989), Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge UK.
54. Hatcher, L., (1994). A step-by-step approach to using the SAS system for factor analysis and structural equation modeling. Cary, NC: SAS Institute. Focus on the CALIS procedure.
55. Hattie, J. (1985) Methodology Review: Assessing unidimensionality of
56. tests and items. Applied Psychological measurement, 9, 2, 139-164.
57. Hollenstein, H. (1996): A Composite Indicator of a Firm's Innovativeness. An Empirical Analysis Based on Survey Data for Swiss Manufacturing, Research Policy, 25, 633-45.
58. Hollenstein, H. (2003): Innovation Modes in the Swiss Service Sector. A Cluster Analysis Based on Firm-level Data, Research Policy, 32(5), 845-863.
59. Homma, T. and Saltelli, A. (1996) Importance measures in global sensitivity analysis of model output. Reliability Engineering and System Safety, 52(1), 1-17.
60. Hutcheson, G., and Sofroniou N.,(1999). The multivariate social scientist: Introductory statistics using generalized linear models. Thousand Oaks, CA: Sage Publications.
61. Jacobs, R. P. Smith and M. Goddard (2004) Measuring performance: an examination of composite performance indicators, Centre for Health Economics, Technical Paper Series 29.
62. Jae-On K., and Mueller C.W. (1978a). Introduction to factor analysis: What it is and how to do it. Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 13.
63. Jae-On K., and Mueller C.W. (1978b). Factor Analysis: Statistical methods and practical issues. Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 14.
64. Jamison, D. and Sandbu, M. (2001) WHO ranking of health system performance. Science, 293, 1595-1596.
65. Jencks, S.F., Huff, E.D. and Cuerdon, T. (2003) Change in the quality of care delivered to Medicare beneficiaries, 1998-1999 to 2000-2001, Journal of the American Medical Association, 289(3): 305-12.
66. Kahma N. (2000), Measuring environmental quality: an index of pollution, Ecological Economics, vol. 35 pp. 191-202.
67. Kahn J.R and Maynard P. (1995) Conjoint Analysis as a Method of Measuring Use and Non-Use Values of Environmental Goods, paper presented at the American Economic Association.
68. Kahn J.R, (1998), Methods for aggregating performance indicators, mimeo, University of Tennessee.
69. Kahnna N., (2000), Measuring Environmental quality: an index of pollution, Ecological Economics, 32, 191-202.

70. Karlsson J. (1998), A systematic approach for prioritizing software requirements, PhD. Dissertation n. 526, Linkoping, Sverige.
71. Kaufmann D., Kraay A., and Zoido-Lobatón P., (1999), Aggregating Governance Indicators, Policy Research Working Papers, World Bank, http://www.worldbank.org/wbi/governance/working_papers.html
72. Kaufmann D., Kraay A., and Zoido-Lobatón P., (2003), Governance matters III: governance Indicators for 1996-2002, mimeo, World Bank.
73. Keeney R., and Raiffa H. (1976) - Decision with multiple objectives: preferences and value trade-offs, Wiley, New York.
74. Keynes, J. M., (1891), *The Scope and Method of Political Economy*. London: Macmillan.
75. King's Fund (2001), The sick list 2000, the NHS from best to worst. http://www.fulcrumtv.com/sick%20list.htm
76. Kline, R.B. (1998). Principles and practice of structural equation modeling. NY: Guilford Press. Covers confirmatory factor analysis using SEM techniques. See esp. Ch. 7.
77. Koedijk K., and Kremers J., (1996), Market opening, regulation and growth in Europe, Economic Policy (0)23, October.
78. Korhonen P., Tainio R., and Wallenius J., (2001), Value efficiency analysis of academic research, European Journal of Operational Research, 130, 121-132.
79. Krantz D.H., Luce R.D., Suppes P. and Tversky A. (1971) - Foundations of measurement, vol. 1, Additive and polynomial representations, Academic Press, New York.
80. Lawley, D. N. and Maxwell A. E. , (1971). Factor analysis as a statistical method. London: Butterworth and Co.
81. Levine, M.S., (1977). Canonical analysis and factor comparison. Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 6.
82. Little R.J.A., and Schenker N., (1994), Missing Data, in Handbook for Statistical Modeling in the Social and Behavioral Sciences (G. Arminger, C.C Clogg, and M.E. Sobel eds.) pp.39-75, New York: Plenum.
83. Little R.J.A (1997) Biostatistical Analysis with Missing Data, in Encyclopedia of Biostatistics (p. Armitage and T. Colton eds.) London: Wiley.
84. Little R.J.A. and Rubin D.B. (2002), Statistical Analysis with Missing Data, Wiley Interscience, J. Wiley &Sons, Hoboken, New Jersey.
85. Mahlberg B. and Obersteiner M., (2001), Remeasuring the HDI by data Envelopment analysis, Interim report IR-01-069, International Institute for Applied System Analysis, Laxenburg, Austria.
86. Manly B., (1994), Multivariate statistical methods, Chapman & Hall, UK.
87. Massam B.H., (2002), quality of life: public planning and private living, Progress in Planning, vol. 58(3), pp. 141-227.
88. Massart, D.L. and Kaufman, L. (1983), The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, New York: John Wiley & Sons, Inc.
89. McDaniel, C., and Gates R., (1998), Contemporary Marketing Research. West Publishing, Cincinnati, OH.
90. Melyn W., and Moesen W.W., (1991), Towards a synthetic indicator of macroeconomic performance: unequal weighting when limited information is available, Public Economic research Paper 17, CES, KU Leuven.
91. Miller, M.B. (1995) Coefficient Alpha: a basic introduction from the perspectives of classical test theory and structural equation modelling. Structural Equation Modelling, 2, 3, 255-273.
92. Milligan, G.W. and Cooper, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," Psychometrika, 50, 159 -179.
93. Moldan, B. and Billharz, S. (1997) Sustainability Indicators: Report of the Project on Indicators of Sustainable Development. SCOPE 58. Chichester and New York: John Wiley & Sons.

94. Muldur U., (2001), Technical annex on structural indicators. Two composite indicators to assess the progress of member States in their transition towards a knowledge based economy, DG RTD, Brussels.
95. Munda G. (1993), Fuzzy information in multicriteria environmental evaluation models, PhD dissertation Vrije Universiteit te Amsterdam.
94. Munda G. (2004) – MCDA and Sustainability Decisions, forthcoming in J. Figueira, S. Greco and M. Ehrgott (eds.) – State of the art of multiple-criteria decision analysis, Kluwer, Dordrecht. Munda G. (1995) - Multicriteria evaluation in a fuzzy environment, Physica-Verlag, Contributions to Economics Series, Heidelberg. Munda, G. and Nardo, M. (2003) On the methodological foundations of composite indicators used for ranking countries. In OECD/JRC Workshop on composite indicators of country performance, Ispra, Italy, May 12, http://webfarm.jrc.cec.eu.int/uasa/evt-OECD-JRC.asp. See also Munda, G., M. Nardo (2003), "On the Construction of Composite Indicators for Ranking Countries", mimeo, Universitat Autonoma de Barcelona.
97. Nanduri M., Nyboer J., and Jaccard M., (2002), Aggregating physical intensity indicators: results of applying the composite indicator approach to the Canadian industrial sector, Energy Policy, 30, 151-162.
98. Nardo, M., S. Tarantola, A. Saltelli, C. Andropoulos, R. Buescher, G. Karageorgos, A. Latvala, and F. Noel (2004) The e-business readiness composite indicator for 2003: a pilot study EUR 21294.
99. Nicoletti G., S. Scarpetta and O. Boylaud, (2000), Summary indicators of product market regulation with an extension to employment protection legislation, Economics department working papers NO. 226, ECO/WKP(99)18. http://www.oecd.org/eco/eco
100. Nilsson, R. (2000), Confidence Indicators and Composite Indicators, CIRET conference, Paris, 10-14 October 2000.
101. NISTEP (National Institute of Science and Technology Policy), (1995) Science and Technology Indicators, NISTEP Report No. 37, Japan.
102. Norman, G. R., and Streiner D. L. , (1994). Biostatistics: The bare essentials. St. Louis, MO: Mosby.
103. Nunnaly, J. (1978). Psychometric theory. New York: McGraw-Hill.
104. OECD (1999) Employment Outlook, Paris.
105. OECD, (2003a), Quality Framework and Guidelines for OECD Statistical Activities, available on www.oecd.org/statistics.
106. OECD (2003b) – Composite indicators of country performance: a critical assessment, DST/IND(2003)5, Paris.
107. Parker J., (1991) Environmental reporting and environmental indices. PhD Dissertation, Cambridge, UK.
108. Pett, M.A., Lackey, N.R. and Sullivan J.J., (2003). Making sense of factor analysis: The use of factor analysis for instrument development in health care research. Thousand Oaks, CA: Sage Publications.
109. Pré Consultants (2000) The Eco-indicator 99. A damage oriented method for life cycle impact assessment. http://www.pre.nl/eco-indicator99/ei99-reports.htm
110. Podinovskii V.V., (1994) -Criteria importance theory, Mathematical Social Sciences, 27, pp. 237 - 252.
111. Porter M. and Stern S. (1999) The new challenge to America's prosperity: finding from the Innovation Index, Council on Competitiveness, Washington D.C.
112. Puolamaa M., Kaplas M., and Reinikainen T., (1996) Index of Environmental Friendliness. A methodological study, Eurostat.
113. Raykov, T. (1998b). Cronbach's Alpha and Reliability of Composite with Interrelated Nonhomogenous Items. Applied Psychological Measurement, 22, 375-385.

114. Rosen R., Life Itself - A Comprehensive Inquiry into Nature, Origin, and Fabrication of Life. Columbia University Press 1991.
115. Roy B. (1996) - Multicriteria methodology for decision analysis, Kluwer, Dordrecht.
116. Saaty, T. L. (1980) The Analytic Hierarchy Process, New York: McGraw-Hill.
117. Saaty, R.W. (1987) The analytic hierarchy process- what it is and how it is used. Mathematical Modelling, **9**, 161-176.
118. Saisana, M. and Tarantola, S. (2002) State-of-the-art report on current methodologies and practices for composite indicator development, EUR 20408 EN, European Commission-JRC: Italy.
119. Saisana M., Tarantola S., Saltelli A. (2005) Uncertainty and sensitivity techniques as tools for the analysis and validation of composite indicators, Journal of the Royal Statistical Society A, 168(2), 1-17.
120. Sajeva, M. (2004) A methodology for quality assurance of knowledge economy statistical indicators, ERU report, in publication.
121. Saltelli, A. (2002) Making best use of model valuations to compute sensitivity indices. Computer Physics Communications, **145**, 280-297.
122. Saltelli, A., Chan, K. and Scott, M. (2000a) Sensitivity analysis, Probability and Statistics series, New York: John Wiley & Sons.
123. Saltelli, A. and Tarantola, S. (2002) On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal, Journal of American Statistical Association, 97 (459), 702-709.
124. Saltelli, A., Tarantola, S. and Campolongo, F. (2000b) Sensitivity analysis as an ingredient of modelling. Statistical Science, **15**, 377-395.
125. Saltelli, A., Tarantola, S., Campolongo, F. and Ratto, M. (2004) Sensitivity Analysis in practice, a guide to assessing scientific models. New York: John Wiley & Sons. A software for sensitivity analysis is available at http://www.jrc.cec.eu.int/uasa/prj-sa-soft.asp.
126. Schumpeter, J.A., (1933), The common sense of econometrics, *Econometrica* 1: 5-12.
127. Sen A. 1989 Development as Capabilities Expansion, Journal of Development Planning vol.19, 41-58
128. Sharpe, A., (2004), *Literature Review of Frameworks for Macro-indicators*, Centre for the Study of Living Standards, Ottawa, CAN.
129. Sobol', I. M. (1993) Sensitivity analysis for non-linear mathematical models. Mathematical Modelling & Computational Experiment **1**, 407-414.
130. Sobol', I. M. (1967) On the distribution of points in a cube and the approximate evaluation of integrals. USSR Computational Mathematics and Physics, **7**, 86–112.
131. Spath, H. (1980), Cluster Analysis Algorithms, Chichester, England: Ellis Horwood.
132. SPRG (2001) Report of the Scientific Peer Review Group on Health Systems Performance Assessment, Scientific Peer Review Group (SPRG), WHO: Geneva. http://www.who.int/health-systemsperformance/sprg/report_of_sprg_on_hspa.htm
133. Storrie D. and Bjurek H., (1999), Benchmarking European labour market performance with efficiency frontier technique. Discussion Paper FS I 00-2011.
134. Storrie D. and Bjurek H., (2000), Benchmarking the basic performance indicators using efficiency frontier techniques, Report presented to the European commission, DG employment and social affairs.
135. Tarantola, S., Jesinghaus, J. and Puolamaa, M. (2000) Global sensitivity analysis: a quality assurance tool in environmental policy modelling. In Sensitivity Analysis (eds A. Saltelli, K. Chan, M. Scott) pp. 385-397. New York: John Wiley & Sons.
136. Tarantola, S., Saisana, M., Saltelli, A., Schmiedel, F. and Leapman, N. (2002) Statistical techniques and participatory approaches for the composition of the European Internal Market Index 1992-2001, EUR 20547 EN, European Commission: JRC-Italy.

137. Ting H.M. (1971), Aggregation of attributes for multiattributed utility assessment, Technical report n. 66, Operations Research Center, MIT Cambridge Mass.
138. Ülengin B., Ülengin F., Güvenç Ü., (2001), A multidimensional approach to urban quality of life: the case of Istanbul. European Journal of Operational Research 130, 361-374.
139. United Nations (1999, 2000, 2001) Human Development Report. United Kingdom: Oxford University Press. http://www.undp.org
140. U.S. Department of Energy and Energy Information Administration, (1995), Measuring energy efficiency un the United States' economy: a beginning. U.S. Department of Energy, Washington, DC.
141. Vichi M., and Kiers, H. (2001) Factorial k-means analysis for two-way data, Computational Statistics and Data Analysis, 37(1), 49-64.
142. Vincke Ph. (1992) - Multicriteria decision aid, Wiley, New York.
143. Widaman, K. F. (1993). Common factor analysis versus principal components analysis: Differential bias in representing model parameters?" Multivariate Behavioral Research 28: 263-311. Cited with regard to preference for PFA over PCA in confirmatory factor analysis in SEM.
144. World Economic Forum (2002) Environmental Sustainability Index http://www.ciesin_org/indicators/ESI/index.html.
145. WHO (2000), Overall Health System attainment. http://www.who.int/whr2001/2001/archives/2000/en/contents.htm
146. Young H.P. and Levenglick A. (1978) A consistent extension of Condorcet's election principle, SIAM Journal of Applied Mathematics, 35, pp. 285-300.
147. Young H.P. (1988) – Condorcet's theory of voting, American Political Science Review, Vol. 82, No. 4, pp. 1231-1244.
148. Zimmermann H.J. and Zysno P. (1983) Decisions and evaluations by hierarchical aggregation of information, Fuzzy Sets and Systems, 10, pp.243-260.

APPENDIX

TAI is made of a relatively small number of sub-indicators, which renders it suitable for the didactic purposes, is well documented by its developers, raw data are freely available on the WEB, and issues of technological development are of importance to society and often discussed on newspapers/press. For explanatory purposes we consider only the first 23 of the 72 original countries.

TAI focuses on four dimensions of technological capacity (Table A.1 in the Appendix):

(a) Creation of technology. Two sub-indicators are used to capture the level of innovation in a society: the number of patents granted per capita (to reflect the current level of invention activities), and the receipts of royalty and license fees from abroad per capita (to reflect the stock of successful innovations of the past that are still useful and hence have market value).
(b) Diffusion of recent innovations. This diffusion is measured by two sub-indicators: diffusion of the Internet (indispensable to participation), and by exports of high-and medium-technology products as a share of all exports.
(c) Diffusion of old innovations. Two sub-indicators are included here, telephones and electricity, which are especially important because they are needed to use newer technologies and are also pervasive inputs to a multitude of human activities. Both indicators are expressed as logarithms, as they are important at the earlier stages of technological advance but not at the most advanced stages. Expressing the measure in logarithms ensures that as the level increases, it contributes less to the technology achievement.
(d) Human skills. A critical mass of skills is indispensable to technological dynamism. The foundations of such ability are basic education to develop cognitive skills and skills in science and mathematics. Two sub-indicators are used to reflect the human skills needed to create and absorb innovations: mean years of schooling and gross enrolment ratio of tertiary students enrolled in science, mathematics and engineering.

Table A.2 in the Appendix shows the raw data for the eight sub-indicators for a set of 72 countries (original). However the original data set contains a large number of missing values, mainly due to missing data in Patents and Royalties. Therefore in the section 4 on imputation, the entire dataset will be used for the estimation of the missing values. For explanatory purposes a set of the 23 countries (from Finland to Slovenia) will be used throughout this document for the presentation of the different methodologies for the analysis and construction of a composite indicator.

Table A.1. List of sub-indicators of the Technology Achievement Index

Indicator	*Unit*	*Definition*
Creation of technology		
PATENTS	Patents granted per 1,000,000 people	Number of patents granted to residents, so as to reflect the current level of invention activities (1998)
ROYALTIES	US $ per 1,000 people	Receipts of royalty and license fees from abroad per capita, so as to reflect the stock of successful innovations of the past that are still useful and hence have market value (1999)
Diffusion of recent innovations		

INTERNET	Internet hosts per 1,000 people	Diffusion of the Internet, which is indispensable to participation in the network age (2000)
EXPORTS	%	Exports of high and medium technology products as a share of total goods exports (1999)
Diffusion of old innovations		
TELEPHONES	Telephone lines per 1,000 people (log)	Number of telephone lines (mainline and cellular), which represents old innovation needed to use newer technologies and is also pervasive input to a multitude of human activities (1999)
ELECTRICITY	kWh per capita (log)	Electricity consumption, which represents old innovation needed to use newer technologies and is also pervasive input to a multitude of human activities (1998)
Human skills		
SCHOOLING	years	Mean years of schooling (age 15 and above), which represents the basic education needed to develop cognitive skills (2000)
ENROLMENT	%	Gross enrolment ratio of tertiary students enrolled in science, mathematics and engineering, which reflects the human skills needed to create and absorb innovations (1995-1997)

Table A.2. Raw data for the sub-indicators of the Technology Achievement Index. The first 23 countries are used as case study in the document. Units are given in Table A.1.

		PATENTS	ROYALTIES	INTERNET	EXPORTS	TELEPHONES (log)	ELECTRICITY (log)	SCHOOLING	ENROLMENT
1	Finland	187	125.6	200.2	50.7	3.08	4.15	10	27.4
2	United States	289	130	179.1	66.2	3.00	4.07	12	13.9
3	Sweden	271	156.6	125.8	59.7	3.10	4.14	11.4	15.3
4	Japan	994	64.6	49	80.8	3.00	3.86	9.5	10
5	Korea, Rep. of	779	9.8	4.8	66.7	2.97	3.65	10.8	23.2
6	Netherlands	189	151.2	136	50.9	3.02	3.77	9.4	9.5
7	United Kingdom	82	134	57.4	61.9	3.02	3.73	9.4	14.9
8	Canada	31	38.6	108	48.7	2.94	4.18	11.6	14.2
9	Australia	75	18.2	125.9	16.2	2.94	3.94	10.9	25.3
10	Singapore	8	25.5	72.3	74.9	2.95	3.83	7.1	24.2
11	Germany	235	36.8	41.2	64.2	2.94	3.75	10.2	14.4
12	Norway	103	20.2	193.6	19	3.12	4.39	11.9	11.2
13	Ireland	106	110.3	48.6	53.6	2.97	3.68	9.4	12.3
14	Belgium	72	73.9	58.9	47.6	2.91	3.86	9.3	13.6
15	New Zealand	103	13	146.7	15.4	2.86	3.91	11.7	13.1
16	Austria	165	14.8	84.2	50.3	2.99	3.79	8.4	13.6
17	France	205	33.6	36.4	58.9	2.97	3.80	7.9	12.6
18	Israel	74	43.6	43.2	45	2.96	3.74	9.6	11
19	Spain	42	8.6	21	53.4	2.86	3.62	7.3	15.6
20	Italy	13	9.8	30.4	51	3.00	3.65	7.2	13
21	Czech Republic	28	4.2	25	51.7	2.75	3.68	9.5	8.2
22	Hungary	26	6.2	21.6	63.5	2.73	3.46	9.1	7.7
23	Slovenia	105	4	20.3	49.5	2.84	3.71	7.1	10.6
24	Hong Kong, China (SAR)	6		33.6	33.6	3.08	3.72	9.4	9.8
25	Slovakia	24	2.7	10.2	48.7	2.68	3.59	9.3	9.5
26	Greece			16.4	17.9	2.92	3.57	8.7	17.2
27	Portugal	6	2.7	17.7	40.7	2.95	3.53	5.9	12
28	Bulgaria	23		3.7	30	2.60	3.50	9.5	10.3
29	Poland	30	0.6	11.4	36.2	2.56	3.39	9.8	6.6
30	Malaysia			2.4	67.4	2.53	3.41	6.8	3.3
31	Croatia	9		6.7	41.7	2.63	3.39	6.3	10.6
32	Mexico	1	0.4	9.2	66.3	2.28	3.18	7.2	5
33	Cyprus			16.9	23	2.87	3.54	9.2	4
34	Argentina	8	0.5	8.7	19	2.51	3.28	8.8	12
35	Romania	71	0.2	2.7	25.3	2.36	3.21	9.5	7.2
36	Costa Rica		0.3	4.1	52.6	2.38	3.16	6.1	5.7
37	Chile		6.6	6.2	6.1	2.55	3.32	7.6	13.2
38	Uruguay	2		19.6	13.3	2.56	3.25	7.6	7.3
39	South Africa		1.7	8.4	30.2	2.43	3.58	6.1	3.4
40	Thailand	1	0.3	1.6	48.9	2.09	3.13	6.5	4.6

41	Trinidad and Tobago			7.7	14.2	2.39	3.54	7.8	3.3
42	Panama			1.9	5.1	2.40	3.08	8.6	8.5
43	Brazil	2	0.8	7.2	32.9	2.38	3.25	4.9	3.4
44	Philippines		0.1	0.4	32.8	1.89	2.65	8.2	5.2
45	China	1	0.1	0.1	39	2.08	2.87	6.4	3.2
46	Bolivia	1	0.2	0.3	26	2.05	2.61	5.6	7.7
47	Colombia	1	0.2	1.9	13.7	2.37	2.94	5.3	5.2
48	Peru		0.2	0.7	2.9	2.03	2.81	7.6	7.5
49	Jamaica		2.4	0.4	1.5	2.41	3.35	5.3	1.6
50	Iran, Islamic Rep. of	1			2	2.12	3.13	5.3	6.5
51	Tunisia		1.1		19.7	1.98	2.92	5	3.8
52	Paraguay		35.3	0.5	2	2.14	2.88	6.2	2.2
53	Ecuador			0.3	3.2	2.09	2.80	6.4	6
54	El Salvador		0.2	0.3	19.2	2.14	2.75	5.2	3.6
55	Dominican Republic			1.7	5.7	2.17	2.80	4.9	5.7
56	Syrian Arab Republic				1.2	2.01	2.92	5.8	4.6
57	Egypt		0.7	0.1	8.8	1.89	2.94	5.5	2.9
58	Algeria				1	1.73	2.75	5.4	6
59	Zimbabwe			0.5	12	1.56	2.95	5.4	1.6
60	Indonesia			0.2	17.9	1.60	2.51	5	3.1
61	Honduras				8.2	1.76	2.65	4.8	3
62	Sri Lanka			0.2	5.2	1.69	2.39	6.9	1.4
63	India	1		0.1	16.6	1.45	2.58	5.1	1.7
64	Nicaragua			0.4	3.6	1.59	2.45	4.6	3.8
65	Pakistan			0.1	7.9	1.38	2.53	3.9	1.4
66	Senegal			0.2	28.5	1.43	2.05	2.6	0.5
67	Ghana				4.1	1.08	2.46	3.9	0.4
68	Kenya			0.2	7.2	1.04	2.11	4.2	0.3
69	Nepal			0.1	1.9	1.08	1.67	2.4	0.7
70	Tanzania, U. Rep. of				6.7	0.78	1.73	2.7	0.2
71	Sudan				0.4	0.95	1.67	2.1	0.7
72	Mozambique				12.2	0.70	1.73	1.1	0.2

Mission of the JRC

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.