Quaderni di Statistica
Vol. 11, 2009

Composite indicators of scientific research

Michela Gnaldi Maria Giovanna Ranalli
Dipartimento di Economia, Finanza e Statistica, Università di Perugia
E-mail: michela.gnaldi@stat.unipg.it, giovanna.ranalli@stat.unipg.it

Summary: The construction of composite measures creates specific methodological challenges. We address these issues through an analysis of some individual indicators put forward by the Italian Steering Committee for Research Evaluation (CIVR). To construct a composite indicator (CI) of scientific research, five normalisation methods, a weighting scheme, and two aggregation schemes have been computed and combined, resulting in 135 CIs. The variation in the rankings assigned by the CIs to the Universities has been explored to gauge the robustness of the CIs rankings. The analysis suggests that the judgements that have to be made in the construction of a CI can have a significant impact on the resulting score and that technical and analytical issues in the design of CIs have important policy implications.

Keywords: University research performance measurement, Composite indicators robustness, Sensitivity Analysis.

1. Introduction

Composite indicators (CIs) integrate a large amount of information in a format that is easily understood and are, therefore, a valuable tool for conveying an overall assessment of performance in priority areas. However, the construction of composite measures creates specific methodological challenges. Any CI may be considered as a model (OECD, 2008; Jacobs *et al.*, 2004; 2006) where the CI is the response variable and the covariates are all the subjective judgements - the sources of uncertainty - which have to be made (e.g. the selection of individual indicators, the choice of normalisation methods, weighting schemes, aggregation model *etc.*). All these potential sources of uncertainty should be addressed because they affect both the variance of the CIs and the variability of any rankings based on CIs. In this context, sensitivity analysis can be considered as an appropriate tool to assess such uncertainty, because it studies how the variation in the output can be apportioned to different sources of variation in the assumptions. Its primary aim is hence to quantify the overall uncertainty in CIs - and in country/institution rankings based on CIs - as a result of the uncertainty in the model inputs.

This work investigates the degree to which composite measures are an appropriate metric for evaluating and ranking the research performance of Italian universities. Do they reflect accurately the performance of universities? To what degree are they influenced by the uncertainty surrounding underlying indicators on which they are based? We address the methodological challenges implied in the construction of CIs through a sensitivity analysis of CIs based on the individual indicators of academic research performance put forward by the Italian Steering Committee for Research Evaluation (CIVR). It is expected that the choice to include a particular indicator in the composite, the choice of a normalisation, aggregation and weighting scheme can have an impact on the rankings of the individual units (e.g. universities) within the composite and that, in a system - such as the Italian one - where universities are rewarded according to the outcome of a CI, greater attention should be paid to the origin and nature of such sources of uncertainty.

2. Dataset description

In 2006 the outcomes of the first large-scale research evaluation process were published in Italy. That exercise was performed by the CIVR and based on the model of the British RAE. Its aim was an assessment of the scientific production of 77 Italian universities (in the period included between 2001 and 2003), 12 public research institutions and 13 private research institutions. The object of the exercise was the evaluation of research products intended in *latu sensu*: not only books, chapters, conference proceedings and scientific articles, but also patents, spin-offs, projects, design and drawings, performances, expositions and exhibitions, manufactured products and works of art. The exercise was performed by panels of experts and each product has been evaluated by at least 2 experts, in terms of their quality, relevance/importance, originality/innovation and international competitiveness. The expert judgment (or score) on each research product has been unique and each product has been judged either excellent, or good, adequate or poor.

The data includes a number of individual indicators (e.g. number of publications, ordinary funds, impact factor of the journal in which each publication is included, number of spin-off created, etc.). Some of these individual indicators are evaluated with reference to university disciplinary areas, whereas others are aggregated by the CIVR at a university level.

An anomaly of the individual indicators put forward by the CIVR is that they depend on the size of the units for which they are calculated (Fabbris *et al.*, 2008). This determines a strong relationship among indicators which end up with representing the same statistical dimension, namely the university size (e.g. big universities have more research products and get more public funds than smaller ones). Such absolute indicators have been therefore transformed into relative indicators.

First, the individual indicators at a level of university disciplinary areas have been aggregated to the university level. Let x_{hqu} be the q-th elementary indicator, for q = 1, …, Q, associated to the h-th disciplinary area (h = 3,…, H = 20) of the u-th university (u = 1, …, U = 77). Then, the university level indicator is given by:

$$x_{qu} = \sum_{h=1}^{H} x_{hqu} w_{hu},$$

with weight w_{hu} given by the quota of professors, lecturers and research fellows in the university disciplinary area h of university u. Then, for each university, standardised individual indicators are calculated as:

$$x^{*}_{qu} = 100 \frac{x_{qu}}{x_{\bar{q}u}}$$

where $x_{\bar{q}u}$ is the indicator average:

$$x_{\bar{q}u} = \sum_{u=1}^{U} x_{qu} w_u$$

and w_u is the quota of professors, lecturers and research fellows coming from university *u*.

Table 1. Individual Indicators of scientific research selected

	Individual Indicators
1	Product score
2	PRIN funded
3	% of excellent products
4	% of product at least good
5	% of product at least appropriate
6	% of products with IF
7	Patents activated abroad
8	Active spin-off
9	Active partnerships
10	Economic Valorisation of Research Products Index
11	Patent score
12	% of Phd and postdoc students
13	Ability to get funds
14	Research Internationalisation

Factor analysis based on such standardised individual indicators showed the existence of two main dimensions of academic research, the one related to the quality of research and the other related to the ability to valorise - in economic terms - the research activity. An exploratory analysis of the data and a correlation matrix highlighted a number of simple indicators significantly correlated to each other. After having excluded some redundant and incongruous individual indicators and the selection of those most meaningful ones, a set of 14 individual indicators has been chosen to construct Composite Indicators (CIs) of scientific research (see Table 1). For each of the simple indicators described below - with the exception of *PRIN funded* and *Patent score* - the computation procedure and the weights applied were defined by the CIVR itself.

In particular, the indicators are:

- *Product score* is a weighted mean of the scores assigned by the experts to the research products, with weight 1 given to products which have been evaluated excellent, 0.8 given to products which have been evaluated good, 0.6 given to products which have been evaluated appropriate, and 0.2 given to products which have been evaluated poor. The indicator takes values between 0 and 1: it takes value 0 when all the submitted products are not assessable, and value 1 when all the products have been evaluated as excellent;

- *PRIN funded is* the average number of research projects of national interest – over the three years considered – funded through public funds. The indicator was taken from the MIUR-Cineca dataset;
- *Percentage of excellent products* is the percentage of excellent products out of the total number of evaluated products;
- *Percentage of at least good products* is the ratio (multiplied by 100) between the sum of excellent and good products and the total number of evaluated products;
- *Percentage of at least appropriate products* is the ratio (multiplied by 100) between the sum of excellent, good and acceptable products and the total number of evaluated products;
- *Percentage of products with Impact Factor* is the ratio (multiplied by 100) between the number of products with IF and the total number of submitted products;
- *Patents activated abroad* is the total number of patents activated abroad over the period 2001-2003;
- *Active spin-off* is the total number of spin-offs activated over the period 2001-2003;
- *Active partnerships* is the total number of partnerships activated over the period 2001-2003;

- *Economic Valorisation of Research Products Index* is obtained as a weighted mean of the number of submitted patents (with weight 1.5 given to those submitted abroad), of the number of active patents (with weight 1.5 given to those activated abroad), of the income deriving from the patent trade, of the number of active spin-offs and of the number of active partnerships. To the five terms, the CIVR applied the following weights: 1, 1, 2, 4 and 2, respectively;
- *Patent score* is an indicator analogous to the *Product Score* indicator that we calculated as a weighted mean of the scores given by the panel of experts to the patents submitted for evaluation, with weight 1 given to patents which have been evaluated excellent, 0.8 given to patents which have been evaluated good, 0.6 given to patents which have been evaluated appropriate, and 0.2 given to patents which have been evaluated poor;
- *Percentage of Phd and postdoc students* is the ratio (multiplied by 100) between the total number of Phd and postdoc students and the total number of professors, assistant professors and lecturers. It is an index which express the capability of a university to attract young researchers;
- *Ability to get funds* is the total number of funding received by the universities through State and EU funding and other international and national funding bodies;
- *Research Internationalisation* is the percentage of professors, assistant professors and lecturers in mobility (e.g. Italian professors working abroad and foreign professors moving to Italy) out of the total number of professors, assistant professors and lecturers.

3. Case study: methodology and results

Any composite indicator (CI) may be considered as a model (OECD, 2008) where the CI is the response variable and the covariates are all the subjective judgements (e.g. the sources of uncertainties) which have to be made (e.g. the selection of individual indicators, the choice of normalisation methods, weighting schemes, aggregation model *etc*.). It is argued (Saisana *et al.*, 2005; Munda *et al.,* 2005, Saltelli, 2007) that all these potential sources of uncertainty should be addressed because they affect both the variance of the CIs and the variability of any ranking based on CIs. Sensitivity analysis can be considered as an appropriate tool to assess such uncertainties. In fact, sensitivity analysis studies how the variation in the output can be apportioned to different sources of variation in the assumptions, and how the given composite indicator depends upon the information fed into it. Given that its primary aim is to quantify the overall uncertainty in country/institution rankings as a result of the uncertainties in the model input, it can help to gauge the robustness of the composite indicator ranking and to identify which countries/institutions are favoured or weakened under certain assumptions.

A regression analysis has been employed to assess the contribution of the individual sources of uncertainty to the variance of the CIs. The sources of uncertainty introduced in the model are: (*i*) inclusion and exclusion of the 14 individual indicators; (*ii*) alternative data normalisation schemes; (*iii*) different aggregation schemes. The model produces estimates for the university average effect, for the normalisation methods, aggregation schemes and single indicator effects, and for possible interactions.

Normalisation is required prior to data aggregation as the indicators in our data have different measurement units. We have considered the following normalisation methods:

1. Ranking (n1):

$$I_{qu} = Rank(x^{*}_{qu})$$

2. Standardisation (n2):

$$I_{qu} = (x^{*}_{qu} - \mu_u(x^{*}_{qu})) / \sigma_u(x^{*}_{qu})$$

3. Min-Max (n3):

$$I_{qu} = \frac{(x^{*}_{qu} - \min_u(x^{*}_{qu}))}{(\max_u(x^{*}_{qu}) - \min_u(x^{*}_{qu}))}$$

4. Distance to a reference university (n4):

$$I_{qu} = \frac{x_{qu}^{*}}{\max_u (x_{qu}^{*})}$$

5. Categorical scales (n5): universities are divided into four groups according to quartiles and I_{qu} takes value in the set [0, 25, 50, 75, 100] according to the group whom the university belongs to.

As of weighting schemes, the existing literature offers a quite rich menu of alternative methods (Melyn *et al.*, 1991; Saaty, 1987). In this preliminary work, weights have been computed by means of Factor analysis and factors have been extracted through Maximum Likelihood, giving weights that intervene to correct for overlapping information between two or more correlated indicators (and not as measures of the theoretical importance of the associated indicator).

The two following aggregation methods have been considered:

1. Linear Aggregation (a1):

$$CI_u = \sum_{q=1}^{Q} w_q I_{qu},$$

2. Geometric Aggregation (a2):

$$CI_u = \prod_{q=1}^{Q} I_{qu}^{w_q}.$$

In both linear and geometric aggregations, weights express trade-offs between indicators (a deficit in one dimension can thus be offset/compensated by a surplus in another); however, in a linear aggregation the compensation is constant, while with geometric aggregations compensation is lower for the CIs with low values.

The five normalisation methods, the weighting scheme, and the two aggregation schemes have been combined, giving 10 combinations and as many CIs and ranking. From the 10 combinations, the one between geometric aggregation and z-scores normalisation has been excluded since z-score normalisation takes negative values. For each of the 9 resulting combinations, the rankings have been calculated on a full model (with all the 14 simple indicators included) and then dropping a simple indicator at a time, resulting in 135 CIs. The variation in the rankings assigned by the CIs to the universities considered has been explored graphically to gauge the robustness of the CIs rankings. Private universities and Special Schools (e.g. Sissa Trieste, Pisa Sant'Anna and Pisa Normale) have been excluded from these analyses. A first output of sensitivity analysis is reported in Figure 1, where universities are ordered by their median rank and the width of the 5th – 95th percentile bounds is also showed.

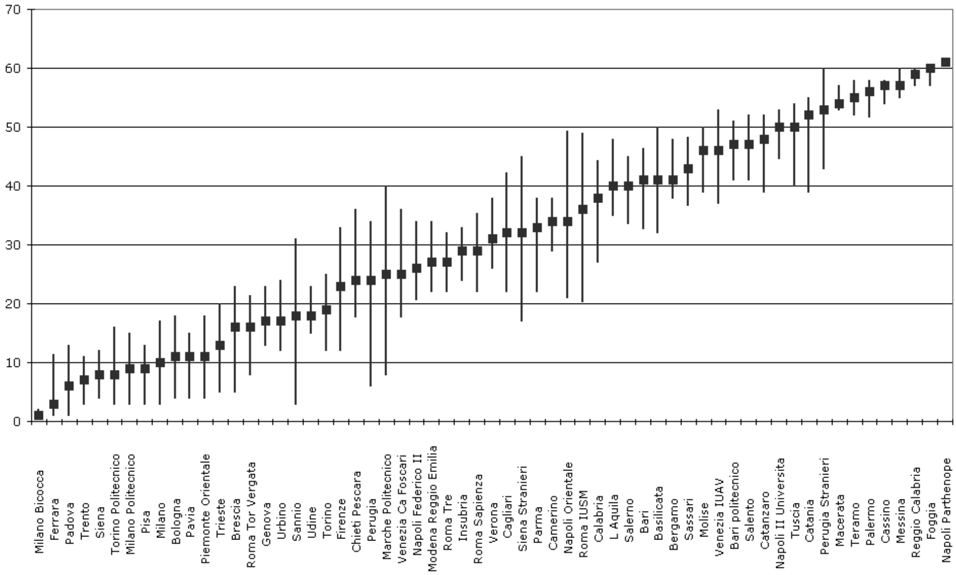

Figure 1. Median CI of universities with 90% interval of variation over the 135 CIs obtained

The width of the 5th – 95th percentile bounds and the ordering of the medians show that Milano Bicocca and Napoli Parthenope are, respectively, stably first and last in the ranking and are not therefore sensible to variations to any of the uncertainty sources of the model. The group of university laggards (and, to some extent, the group of university leaders) is less sensitive to variations, while the central ranking positions are occupied by the most sensible universities. These latter universities are Perugia, Sannio, Siena Stranieri, Napoli Orientale, Roma IUSM and Marche Politecnico, with a variation in their ranking of around 30 positions.

A regression analysis has been employed to assess the contribution of the individual sources of uncertainty to the variance of the CIs. The model produces estimates for the university average effect, for the normalisation methods, aggregation schemes and single indicator effects, and for the interaction effect. The model estimates highlighted that there is a strong university effect on the ranking variability; however, some sources of uncertainty strongly affect the ranking variability. In particular:

i. there is a significant effect of all normalisation methods, but n5. This is because this last normalisation scheme, like n1 and differently from the other normalisation schemes, are not affected by outliers;

ii. of the 14 individual indicators considered, the ones that most affect the ranking variability are the *Product score* and the *Patent Score;*

iii. the individual indicators that affect the least the ranking variability are the *PRIN funded*, the *Research Internationalisation*, the *Percentage of product with IF*, the *Ability to get funds;*

iv. the interaction between n5 and a2 is a strong source of uncertainty of the model. The interaction has mainly a negative significant effect on some university ranking and is one of the most important source of ranking variation for the most sensitive universities.

A regression analysis run per university allows to find out which sources of uncertainty mostly affect the ranking variability of each university, shedding light on the contribution of the single factors on the overall performance of a given university. This analysis is particularly informative when focused on the universities with the highest ranking variability; the model for the most variable university (Marche Politecnico, see Table 2) highlights a conflicting behaviour on the side of the research quality on one hand, and on the side of the ability to valorise, in economic terms, the research activity, on the other hand. Among the single indicators, the most influent sources of uncertainty are *Product Score* and *Percentage of excellent products* – which determine a positive shift in its mean ranking of, respectively, six and five positions – and *Active spin-off* and *Patent score* – which determine a negative shift of, respectively, eleven and fourteen positions. As a consequence of these disagreeing performances on different single indicators, the normalisation schemes – *Standardisation*, *Min-Max* and *Distance to a reference University* – all turn up to be very important sources of uncertainty of the model.

Figure 2 shows the percentage of total variance explained by each of the model uncertainties for each of the 61 universities considered. The most important source of uncertainty is the normalisation scheme, followed by the selection of the simple indicator, the interaction terms and the aggregation schemes. The normalisation schemes are also the source of uncertainty most variable between universities, explaining from a minimum of 4% to a maximum of 85% of the total variance of the model.

4. Concluding remarks

Composite indices are a useful communication and political tool for conveying summary performance information in a relatively simple way. They are used widely in various sectors in public services and are currently used to allocate the 7% quota of public funding to the Italian universities. Composite performance indicators have a number of pros, such as offering a more rounded assessment of performance and presenting the ‘big picture’ in a way in which the public can understand.

Table 2. Regression Analysis for Marche Politecnico

Variable Names	**Estimate**	**Std Error**	**t value**	**Pr(>|t|)**
Intercept	33.541	1.404	23.885	<0.001
Normalisations (Ranking is the reference category)				
Standardisation	-11.867	1.242	-9.552	<0.001
Min-Max	-16.200	1.242	-13.040	<0.001
Distance to a reference university	-22.067	1.242	-17.763	<0.001
Categorical Scales	-1.533	1.242	-1.234	0.220
Aggregation (Linear Aggregation is the reference category)				
Geometric Aggregation	3.000	1.242	2.415	0.017
Exclusion of indicators (all indicators is the reference category)				
Product score	-6.222	1.604	-3.880	<0.001
PRIN funded	0.444	1.604	0.277	0.782
% of excellent products	-5.444	1.604	-3.395	0.001
% of product at least good	-1.889	1.604	-1.178	0.241
% of product at least appropriate	-3.444	1.604	-2.148	0.034
% of products with IF	1.000	1.604	0.624	0.534
Patents activated abroad	0.778	1.604	0.485	0.629
Active spin-off	11.000	1.604	6.859	<0.001
Active partnerships	-0.222	1.604	-0.139	0.890
Economic Valorisation of Research Products Index	0.111	1.604	0.069	0.945
Patent score	14.000	1.604	8.729	<0.001
% of Phd and postdoc students	-1.889	1.604	-1.178	0.241
Ability to get funds	-0.111	1.604	-0.069	0.945
Research Internationalisation	-0.222	1.604	-0.139	0.890
Iteractions				
Min-Max*Geometric aggregation	-4.933	1.757	-2.808	0.006
Distance to a reference university*Geometric aggregation	-4.000	1.757	-2.277	0.025
Categorical Scales*Geometric aggregation	3.533	1.757	2.011	0.047

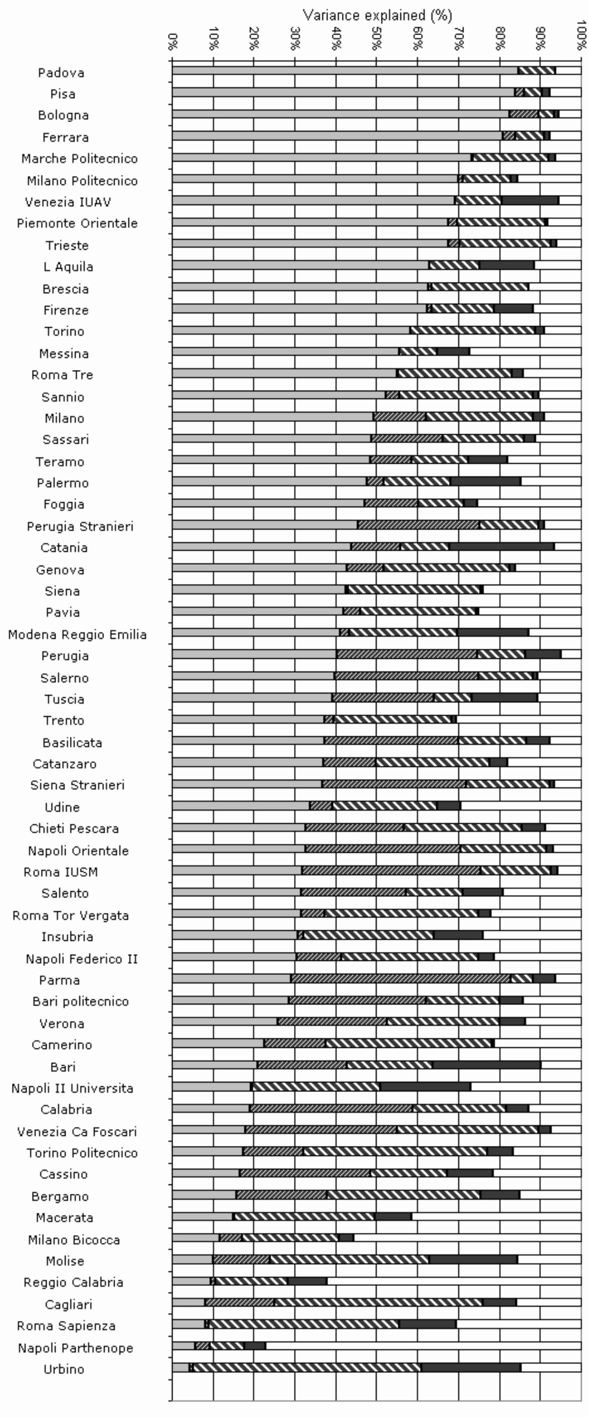

Figure 2. Percentage of variance explained by the model uncertainties per university

It is likely therefore that they will continue to be used in the future in many policy areas. However, in the construction of composite indicators many methodological issues need to be addressed carefully if the results are not to be misinterpreted and manipulated. These results have important scientific, policy and practice implications.

The contribution of this work is an illustrations of the impact of uncertainty on performance measures of the Italian universities. We showed that:

- The university ranking distribution based on CIs is associated with a certain level of variability, which is very strong for some University. In particular, we saw that the groups of university laggards is less sensitive to variations, while the central ranking positions are occupied by the most sensible universities, with a variation in their ranking of around 30 positions (out of 61 positions). As a consequence, it can be argued that dealing with uncertainty in composite performance measures is essential. Any composite indicator of scientific performance needs to be published with indications of uncertainty to communicate the sensitivity of the reported measures. In any performance benchmarking system, it is crucial to know an estimate of the degree of variation for each indicator so that definitive conclusions can be drawn about genuine differences in performances among universities. It is well known (Hicks, 2009), that the use of composite performance measures can generate both positive and negative behavioural responses: therefore, if significant policy and practice decisions rest on the outcome of the CI, it is important to have an understanding of the risks involved in constructing a composite and a ranking.

- It has been illustrated that a very important source of uncertainty is the normalisation scheme, together with the selection of the simple indicator. Besides, the normalisation schemes used are the source of uncertainty most variable between universities. Therefore, the choice of a normalisation, aggregation and weighting scheme has a significant impact on the rankings of individual units within the composite. Great attention should be paid to the origin and nature of such sources of uncertainty because subtle changes to them can noticeably impact on the composite index and rankings of universities.

- Of the 14 individual indicators considered, the ones which mostly affect the ranking variability are the *Product score* and the *Patent Score,* while the individual indicators which less affect the ranking variability are the *PRIN funded*, the *Research Internationalisation*, the *Percentage of product with IF* and the *Ability to get funds.* The regression analysis run per university revealed that the same single indicator can be, at the same time, a strong source of strength and weakness for different universities, determining a considerable positive or negative variation in their ranking positions. Hence, the choice to include a single indicator in the composite is, again, a crucial phase in the construction of a composite measure and a political choice which strongly affects the composite and the ranking. In a system where universities are rewarded according to the outcome of the composite indicator, these decision rules need to be treated with caution and to be publicly defined, not only to make the whole process more transparent *per se*, but also to make the results more acceptable and the reward/penalty system more acceptable in turn.

5. Future directions for research

In this work weights have been computed by means of Factor Analysis, giving weights that intervene to correct for overlapping information between two or more correlated indicators, and not as measures of the theoretical importance of the associated indicator. However, as in common practice, a larger weight could be given to components which are considered more significant in the context of a particular composite indicator. Other weighting schemes - such as those based on experts' judgments of the relative importance of an indicator, or the Data Envelopment Analysis (Moesen *et al.*, 2008) which estimates an efficiency frontier to use as a benchmark to measure the relative performance of countries/institutions - should be accounted for in the next steps of our work.

Besides, when a number of variables are used to evaluate a set of institutions, some may be in favour of one particular institution, while others will favour another. As a consequence, a conflict among the variables could arise. This conflict can be treated in the light of a non-compensatory logic by utilising a discrete non-compensatory multi-criteria approach. Differently from the geometric or linear aggregations already implemented, this approach should take into account the absence of preferential independence and assure non-compensation.

Acknowledgments: Work supported by the project PRIN 2007 "Models and methods for the evaluation of the social and professional return of the competences formed in a university study programme and analysis of the Italian study programmes in statistics and medicine".

References

Fabbris L., Gnaldi M. (2008), Indicatori di valutazione della qualità della ricerca negli atenei: sensibilità, sostituibilità e capacità discriminatoria, in *Professionalità nei servizi innovativi per studenti universitari, (Eds:* Fabbris L., Boccuzzo G. and Martini M.C.), 139–171, CLEUP, Padova.

Hicks D. (2009), Evolving regimes of multi-university research evaluation, *Higher Education*, 57*,* 393-404.

Jacobs R., Smith P., Goddard M. (2004), Measuring performance: an examination of composite performance indicators, *Technical Paper Series* 29, Centre for Health Economics, University of York.

Jacobs R., Goddard M., Smith P.C. (2006), Public Services: Are Composite Measures a Robust Reflection of Performance in the Public Sector, *CHE Research Paper 16,* Centre for Health Economics, University of York.

Melyn W., 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.

Moesen L., Rogge W., Van Puyenbroeck N.T., Saisana M., Saltelli A., Liska R., Tarantola S. (2008), Creating composite indicators with DEA and robustness analysis: the case of the Technology Achievement Index, *Journal of the Operational Research Society,* 59, 239–251.

Munda G., Nardo M. (2005), Constructing Consistent Composite Indicators: the Issue of Weights, EUR 21834 EN, *Joint Research Centre*, Ispra.

OECD (2008), *Handbook on Constructing CIs – Methodology and user guide,* http://composite-indicators.jrc.ec.europa.eu/Handook.htm

Saisana M., Tarantola S., Saltelli A. (2005), Uncertainty and sensitivity techniques as tools for the analysis and validation of CIs, *Journal of the Royal Statistical Society Series* A, 168, 307–323.

Saltelli A. (2007), Composite indicators between analysis and advocacy, *Social Indicators Research*, 81, 65-77.

Saaty R.W. (1987), The analytic hierarchy process: what it is and how it is used, *Mathematical Modelling*, 9, 161–176.