On the “multi-dimensionality” of ranking and the role of bibliometrics in university assessment

WOLFGANG GLÄNZEL AND KOENRAAD DEBACKERE

Steunpunt O&O Indicatoren van de Vlaamse Gemeenschap, KU Leuven, Faculty FBE, Dept. MSI, Belgium

ABSTRACT. The complexity of university activities does not allow reducing the multidimensional space of those activities and their outcomes into one dimension of linear ranking. The difficulty of quantification as well as the all too frequently experienced arbitrariness in defining composite indicators often results in inadequate representation and irreproducibility and hence in clear conflicts with the Berlin Principles on Ranking of HEIs. Even focussing on one single, however important aspect, such as the assessment of research performance, remains a multifaceted endeavour. Using the example of bibliometrics, we point to caveats and pitfalls in the challenge of comparative research assessment of colleges and universities.

KEYWORDS. Bibliometrics, research evaluation, composite indicators, university ranking.

1. Introduction

Performance-based listing of research and education, and above all, the academic ranking of college and universities, has become one of the most favourite issues in the assessment of higher education institutions. At least since the publication of the first edition of the Shanghai Jiao Tong world university ranking in 2003 (ARWU, 2007) and the successor lists, such as the Times Higher Education Supplement – QS World University Rankings in 2005 (THES-QS, 2007), the comparative evaluation of the quality of Education Institutions (HEIs) has been brought into the focus of public and policy interest. World rankings have been followed by national lists in several European countries, in Canada and the US. Although their methodology has been improved since and guidelines for quality management (cf. Berlin Principles on ranking of higher education institutions compiled by the International Ranking Expert Group [IREG, 2006]) have been elaborated, university ranking remains controversial. Methodological and general issues like the questions of how complex multidimensional criteria can be transformed into linearity have been addressed, and are at present vividly discussed. In general parlance, the complexity of university activities does not allow reducing the multidimensional space into one dimension of linear ranking. The difficulty of quantification as well as the all too frequently experienced arbitrariness in defining composite indicators used to result in inadequate representation and irreproducibility and hence in clear conflicts with the Berlin Principles on Ranking of HEIs. Even focussing on one single, however important aspect, such as the assessment of research performance, remains a multifaceted endeavour. Proceeding from our experiences, we illustrate this with two examples. The first one describes the clustering of research institutions on the basis of their publication profiles for comparison of institutional research performance among likes and therefore to avoid the effect of ‘comparing apples with oranges’”. The second example visualises a ‘two-dimensional approach’ to university ranking proceeding from second-generation relational charts. Based on these examples, the study also points to caveats and pitfalls in the challenge of comparative research assessment of universities.

2. A concise discourse on ranking?

Before we tackle the question to which extent reliable and reproducible ranking lists are at all possible, we attempt to clarify the notion of ranking by presenting the following comprehensible but nonetheless precise definition. In verbal terms, ranking is positioning comparable objects on an ordinal scale based on a (non-strict) weak order relation among (statistical) functions of, or a combination of functions of measures or scores associated with those objects.

These (mainly statistical) functions, which are usually based on variables for evaluative purposes, are usually called indicators. Different indicators \( X_i \) representing different aspects of quality, form the components of a composite indicator \( Y \), being the basis of the ranking, particularly,

\[
Y = \sum \lambda_i X_i,
\]

where \( \lambda_i \) (\( i = 1, 2, \ldots, n \)) are \( n \) pre-defined weights and, without loss of generality, we can assume that \( \sum \lambda_i = 1 \). The use of composite indicators always reflects a certain arbitrariness and a level of simplification as we will show in the following. The most problematic issues in applying composite indicators are listed below.

- **Possible interdependence of components**
  The underlying variables represent factors influencing performance. These factors are often not separable and individual variables do consequently not amount to one unique factor each. Variables are therefore often interdependent. For instance, the variables funding, personnel, publication output, citation impact, peer reviews are not independent. The change of one variable can therefore have unpredictable effect upon other variables defining the composite indicator.

- **Altering weights can result in different ranking**
  The choice of weights is in practice arbitrary. The selection is guided by the rankers' preferences rather than by methodological or empirical findings. Rounding to "plausible" values, e.g. 10% or 25%, yet emphasises this arbitrariness.

- **Results might be obscure and irreproducible**
  The mixture of possibly incommensurable indicators, superposition of interdependent variables and arbitrary weighting can make the methodology obscure and the results obtained irreproducible.

- **Random errors of statistical functions are usually ignored**
  Composite indicators are linear combinations of statistical functions which can be subject to random errors themselves. The standard errors of the statistics - such as means and shares - are influenced by the size set of objects measured by the variable in question and the underlying probability distributions. Different positions in the ranking list might therefore be interpreted as ties.

- **The multi-dimensional space is crashed into linearity**
  This is one of the most crucial issues in ranking. From the mathematical viewpoint, a linear combination as applied through the composite indicator is a result of a projection into a subspace. Since projections are irreversible, valuable information is definitely lost by crashing the multi-dimensional space into linearity.

Besides the aforementioned statistical and methodological problems, several data related issues are relevant as well. In the first place, we mention the "cleaness", compatibility and hence the reliability of the data used. Data collection for large-scale ranking still remains a challenge if it is at all feasible. The time-variant nature of the underlying data sources is a
further problem. Thus the incorrect institutional assignment of staff or research-output data taken from different sources might result in incompatibility issues. Combining alumni and staff winning Nobel Prizes and Fields Medals with recent publication- and citation-related data might serve just as an example for such a problematic methodological approach.

3. Selective vs. integrated ranking

In order to account for the complexity of university activities, two basic approaches are possible: the selective and the integrated ranking. The selective ranking focuses on measuring and ranking according to one selected activity whereas the integrated or ‘holistic’ ranking procedure attempts to capture the complex set of all or at least of the most important activities. The advantages of the first method are obvious. As compared with the holistic approach, information loss and incommensurability can be reduced and reliability can be increased. Of course, individual lists have to be prepared for each activity aspect on the other hand. In the following we give a concise description of examples for selective and integrated college and university ranking.

Evaluation of education

In 1993 a national education-related university ranking was published in Germany (Spiegel, 1993). The ranking was survey-based. Questionnaires had been sent to students and professors. A breakdown by fields was presented as well to give a more differentiated picture, to reveal “strengths and weaknesses” and to help students and academic staff make a selection. Because of differences and peculiarities of national educational and accreditation systems, such endeavours are practically restricted to the national level.

Research performance

With the ‘Shanghai Ranking’ (ARWU, 2007), first published in 2003, the focus was shifted to research assessment. The composite indicator build by Shanghai Jiao Tong University is used to rank the world’s major institutes of higher education on the basis of the following weighted key indicators, alumni winning Nobel Prizes and Fields Medals (10%), staff winning Nobel Prizes and Fields Medals (20%), highly-cited researchers according to highlycited.com (20%), articles published in Nature and Science (20%), publications indexed in the Science Citation Index – Expanded (SCIE) and the Social Sciences Citation Index (SSCI) of Thomson Scientific (20%) and the size of the institution (10%). This world-wide ranking was to a large extent facilitated by the availability of the multidisciplinary bibliographic database Web of Science and its derivatives.

‘Holistic approach’

The broader approach chosen by THES-QS, which is largely relying on peer review score, could not overcome the limitations of previous attempts and remained controversial as well. It actually marks a new direction in university ranking, particularly, the trend towards integrated evaluation. The holistic approach, i.e., the comprehensive and integrated quantification of university performance and a world-wide ranking based on all HEI activities, including education, research and third mission, however, remains at least for the present utopian.

The question arises whether there is really any need for an integrated ranking. The evaluation of selected activities within the HEI missions (like quality of education, research performance or the assessment of important third-stream activities) might provide more valuable information for the interested users in the relevant sectors and domains.

The Centre for Higher Education Development (CHE, 2007) has chosen a third way. Their approach is strictly subject oriented but the evaluation extends to both research and education.
The ranking is based on bibliometrics and questionnaires. Although CHE aims at internationalisation, its methodology remains subject to the above-mentioned limitations.

4. Bibliometrics and the “multi-dimensionality” of research activity

In this section, we have a critical look at the possible role of bibliometrics in (selective) university ranking. Although measuring only one, however, important part of research activities, bibliometrics proved an efficient tool in research assessment. Figure 1 sketches the position and function of bibliometrics in quantifying and measuring activities of higher education institutes.

As in the case of all HEIs rankings, first and foremost the following two issues have to be solved for the bibliometric approach as well. The quality of results stands and falls with the correctness of data collection, pre-processing data and the application of sound methodology. This includes correct institutional assignment and the selection of normalised standard indicators that guarantee the robustness and the reproducibility of results.

Another issue arises from the institute-specific specialisation or diversification as even multi-disciplinary research and education institutions have usually more specific research profiles. Thus the practice in institutional evaluation is benchmarking and comparison of institutional performance with reference institutions with similar research profiles. Computerised or semi-computerised classification of research institutions according to their publication profiles (e.g., Thijs and Glänzel, 2008, 2009) can assist both the selection of reference units and the realisation of comparative analysis. From the perspective of validity, comparison of institutions with completely different mission and research profiles should of course be avoided. Although the quantification of research output would in principle allow such treatment, putting business schools and medical schools on the same list would practically not make sense. On the other hand, large universities with originally different profiles like medical and technical universities do have overlapping research activities. Thus the methodology applied should nevertheless be suited for intra- and inter-class comparison where and whenever this makes sense. Finally, an efficient method to further compensate for the biases caused by subject-specific profile heterogeneity in the context of specialisation and diversification, is the consequent standardisation and normalisation of the bibliometric indicators to eliminate subject-specific biases.

In order to obtain a more realistic and differentiated picture of research at higher-educational institutions, the following three scenarios are suggested.

I. Clustering of similar objects
II. Breakdown by fields
III. Standardisation of indicators
The proposed issues are preconditions for the correct use and interpretation of bibliometrics-based indicators and should therefore best be applied in combination.

Clustering of similar objects

Ranking HEIs with completely different profiles as, for instance, based on the comparison of medical schools with business schools, still remains an exercise of “comparing apples with oranges”. In order to find an appropriate profile classification for universities, colleges and research institutes, we have clustered more than 2000 institutions from fifteen European countries according to their publication profiles in the period 2001-2003. The stopping rule introduced by Duda and Hart (1973) was applied to determine and to optimise the number of clusters. The optimum has been found at eight profile clusters. Table 1 presents the classification of European institutions according to Thijs and Glänzel (2008, 2009).

Table 1 The eight clusters resulting from the optimum solution

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1 (Biology)</td>
<td>BIO</td>
</tr>
<tr>
<td>Cluster 2 (Agriculture)</td>
<td>AGR</td>
</tr>
<tr>
<td>Cluster 3 (Multidisciplinary)</td>
<td>MDS</td>
</tr>
<tr>
<td>Cluster 4 (Geo &amp; Space Science)</td>
<td>GSS</td>
</tr>
<tr>
<td>Cluster 5 (Technical &amp; Natural Sciences)</td>
<td>TNS</td>
</tr>
<tr>
<td>Cluster 6 (Chemistry)</td>
<td>CHE</td>
</tr>
<tr>
<td>Cluster 7 (General &amp; Research Medicine)</td>
<td>GRM</td>
</tr>
<tr>
<td>Cluster 8 (Specialised Medicine)</td>
<td>SPM</td>
</tr>
</tbody>
</table>


The application of the university classification alone seems in practice not to be sufficient. Using the example of Belgium, Finland and Spain, one can easily see that national characteristics of scientific research in higher education might strongly influence the constitution of clusters in different countries (cf. Figure 2). Ranking according to clusters would therefore be biased by national representation. Therefore we suggest the additional application of one of both of the following scenarios.

Figure 2 Examples for different national cluster profiles

---

2 This set comprises fifteen countries, namely Switzerland and the members of the European Union before 2004 (EU15) except Greece.
Breakdown by fields

Research performance of a university might differ among their faculties, departments and thus in different science fields. Institute-specific specialisation is often contrasted by, or even combined with, diversification. Research in the same field done by different institutions can still have different profiles as shown in Figure 3. The profile of chemistry research in multidisciplinary universities significantly deviates from that in technical universities, although the overlap is, of course, considerable. In particular, 18% of the publications in chemistry research coming from multidisciplinary universities belong to the organic and medicinal chemistry subfield; the corresponding percentage of the technical universities is equal to 6%. On the other hand, the part of publications classified in the materials science subfield is much higher for the technical than for the multidisciplinary universities (40% for the technical universities vs. 23% for the multidisciplinary universities).

The consequences of these institutional peculiarities are obvious: overall gross publication and citation counts can be misleading when ranking institutions with multidisciplinary and specialised research profiles. This effect can be reduced by breaking down institutional research activity by fields and subfields. This breakdown might further help reveal institutional 'strengths and weaknesses'.

<table>
<thead>
<tr>
<th>Subfield</th>
<th>MDS</th>
<th>TNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>25%</td>
<td>18%</td>
</tr>
<tr>
<td>C2</td>
<td>10%</td>
<td>12%</td>
</tr>
<tr>
<td>C3</td>
<td>18%</td>
<td>0%</td>
</tr>
<tr>
<td>C4</td>
<td>23%</td>
<td>24%</td>
</tr>
<tr>
<td>C5</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>C6</td>
<td>23%</td>
<td>40%</td>
</tr>
</tbody>
</table>


Figure 3 Example of the deviating field structure in different clusters

Standardisation of indicators

An effective method to further compensate for biases caused by subject-specific profile heterogeneity in the context of specialisation and diversification, is the consequent standardisation and normalisation of the bibliometric indicators used in the comparative studies. In an earlier study (Glänzel et al. 2009), we have described an appropriate set of adjusted standard indicators that meets the requirements of meso-level analyses. In particular, proceeding from the indicators used in Budapest and Leuven, we have defined an adequate level of standardisation that makes it possible to use standard indicators for both intra- and inter-cluster analysis, for domain specific as well as multidisciplinary studies and their adequate graphical presentation. The relative indicators developed in Budapest in the 1980s and preferably presented in relational charts (e.g. Schubert and Braun, 1986), have been used as the starting point for the development of the instruments for cross-institutional comparisons. While in the relational charts the Mean Observed Citation Rate (MOCR) was plotted against the journals-based Mean Expected Citation Rate (MECR), the new version of relational charts uses subfield normalised observed and expected citation rates to avoid possible biases caused by subject-specific peculiarities or by different activity profiles as described above. Subject-normalisation is done by dividing the two indicators by the corresponding values of the subject-based Field Expected Citation Rate (FECR). A detailed description of these indicators can be found in (Glänzel et al. 2009).

The subfield abbreviations according to the Leuven subject-classification scheme (Glänzel and Schubert, 2003) are C1 - analytical, inorganic & nuclear chemistry, C2 - applied chemistry & chemical engineering, C3 - organic & medicinal chemistry, C4 - physical chemistry, C5 - polymer science, C6 - materials science.

MOCR is defined as the ratio of citation count to publication count.

The journal-based expected citation rate of a single paper is defined as the average citation rate of all papers published in the same journal.

Analogously to the MOCR, the FECR of a single paper is defined as the average citation rate of all papers published in the same subject in the same year.
While the y-axis presents the factual ‘performance’ measured through citations, the x-axis stands for the impact standard of the journals in which the institution publishes (see Figure 4). Both measures shed light on two important aspects of research assessment, in particular ‘visibility’ and ‘quality-related’ issues. In order to overcome profile-specific biases (which might occur even within the same profile cluster) a strict standardisation and normalisation of indicators should be applied. The two straight lines in the lower right-hand chart of Figure 4, \(x = 1\) and \(y = 1\), indicate two equilibrium situations, particularly, the conformity with the corresponding underlying reference standards (Braun and Glänzel, 1990). A further relative indicator, called Relative Citation Rate (RCR), is obtained as the ratio of the two indicators, where the straight line \(y = x\) indicates the balance between observation and expectation.

The effect of normalisation is demonstrated using the following example. The observed citation impact of 39 European universities in thirteen selected countries (including the three largest HEIs each of the corresponding country) is plotted against its expectation, once without and another time with subject normalisation. The publication period is 2001-2003 and a three-year citation window for each publication year has been applied. The changing ‘ranks’ of medical and technical universities in the two-dimensional presentation are quite impressive.

The scatter plot of (MECR, MOCR) values is presented first in a traditional relational chart (see upper left-hand corner of Figure 4). Both expectation and observation cover quite a large range of citation impact. The fact that SK is a medical university whereas KT is a technical university, certainly contributes to the huge deviations of their respective citation impact indicators. We also mention that both impact indicators (MECR and MOCR) of SK considerably exceed those of the technical university KT. We observe a similar situation for HM and HT, however, at a much lower level. Also HT, as a technical university, appears in the low-end group of this diagram. We obtain a completely different situation if subfield-based normalisation is applied.

The plot of subfield-normalised observation against subfield-normalised expectation can be found in the lower right-hand corner of Figure 4. The positions of universities SK and KT have interchanged; the same applies to HM and HT. The effect of field-specific lower impact of technical universities and the usual high impact of medical of medical universities has thus been eliminated.

Source: Glänzel et al. (2009) based on WoS (Thomson Scientific)

Figure 4 Traditional and subfield-normalised relational chart for 39 selected universities
Two lessons can be learned from this example. (1) Non-normalised, non-standardised counts, such as gross-publication or gross-citation counts, can be strongly affected by university-specific profiles. This effect is measurable even if shares or mean values are used. Only appropriate normalisation can compensate and (nearly) eliminate profile-specific biases. (2) Reducing dimensions means losing information, and might hence result in misinterpretations. The universities AG and NL have almost the same RCR value although both institutions hold different positions in the two charts (cf. Figure 4). While both expected and observed citation impact of AG are in line with the world standard, the indicator values of NL reveal that this university belongs to the high-end with respect to research performance.

Conclusions

The idea of ranking HEIs according to simple, seemingly objective and robust indicators is perhaps tempting. However, robustness is easily lost by building composite indicators with partially interdependent or even incompatible components and arbitrary weights. Reality is more complex than can be described this way. Instead of any linear ranking of colleges and universities, a more detailed, complex analysis is necessary to capture and to reflect several important aspects of performance among the manifold of a university's activities.

Bibliometrics can contribute to evaluate at least one of these aspects. One lesson from bibliometrics is that standardisation and normalisation help eliminate biases and facilitate longitudinal ranking analysis as well. Another lesson from bibliometrics is that even normalisation of indicators cannot disguise that comparing HEIs with completely different profiles nonetheless remains an exercise of 'comparing apples with oranges'.

References

CHE (2007), CHE University Ranking, accessible via: http://www.cheranking.de
W. GLÄNZEL, A. SCHUBERT (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes, Scientometrics, 56 (3), 357-367.
A. SCHUBERT, T. BRAUN (1986), Relative indicators and relational charts for comparative-assessment of publication output and citation impact, Scientometrics, 9 (5-6), 281-291.
B. THIJS, W. GLÄNZEL (2008), A structural analysis of publication profiles for the classification of European research institutes, Scientometrics, 74 (2), 223-236.
B. THIJS, W. GLÄNZEL (2009), A structural analysis of benchmarks on different bibliometrical indicators for European research institutes based on their research profile, Scientometrics, 79 (2), in press. DOI: 10.1007/s11192-009-0425-z.