Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 78, No. 1 (2009) 165–188
DOI: 10.1007/s11192-008-2109-5

Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance

WOLFGANG GLÄNZEL,[a,b] BART THIJS,[a] ANDRÁS SCHUBERT,[b]
KOENRAAD DEBACKERE[a]

[a] *Katholieke Universiteit Leuven, Steunpunt O&O Indicatoren, Dept. MSI, Leuven, Belgium*
[b] *Institute for Research Policy Studies, Hungarian Academy of Sciences, Budapest, Hungary*

A common problem in comparative bibliometric studies at the meso and micro level is the differentiation and specialisation of research profiles of the objects of analysis at lower levels of aggregation. Already the institutional level requires the application of more sophisticated techniques than customary in evaluation of national research performance. In this study institutional profile clusters are used to examine which level of the hierarchical subject-classification should preferably be used to build subject-normalised citation indicators. It is shown that a set of properly normalised indicators can serve as a basis of comparative assessment within and even among different clusters, provided that their profiles still overlap and such comparison is thus meaningful. On the basis of 24 selected European universities, a new version of relational charts is presented for the comparative assessment of citation impact.

Introduction

The mapping and the evaluation of the research performance of institutions and research teams has become one of the principal tasks of present-day bibliometrics. The bibliometric toolbox comprising the appropriate standard indicators has been developed very early in our field but the correct application of these tools to different levels of aggregation is still a challenge. Thus bibliometric meso- and micro-level studies need different standards than, for instance, necessary for comparative macro studies.

Received April 23, 2008

Address for correspondence:
WOLFGANG GLÄNZEL
E-mail: wolfgang.glanzel@econ.kuleuven.ac.be

0138–9130/US $ 20.00

These requirements include issues like statistical reliability based on the lower publication counts, advanced, more fine grained, techniques of data-cleaning, of address and author identification and of reliable subject delineation. Another issue arises from the institute- or group-specific specialisation or diversification. While medium-sized and big countries are practically active in all fields of the sciences and, although one can distinguish at least four different paradigmatic types of national research profiles (cf. [REIST-2, 1997]), their activity patterns still follow the global one to sufficient extent, scientific institutions and research groups have usually more specific research profiles. In contrast to the case of national research performance, 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 & GLÄNZEL, 2008]) can assist both the selection of reference units and the realisation of comparative analysis. An effective method to further compensate 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 the present paper, we will search for an appropriate set of adjusted standard indicators that will meet the requirements of tasks of meso-level analysis, but will also be applicable at other levels of aggregation. We will proceed from the indicators used in Budapest and Leuven to define 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 at ISSRU (Budapest) in the 1980s and preferably presented in *relational charts* (e.g., [SCHUBERT & BRAUN, 1986]), proved useful instruments in cross-national comparisons of national research performance. However, in cross-institutional comparisons, where even multidisciplinary profiles do not necessarily cover all science areas, or may cover most areas not to the same extent, these relational charts might not reflect citation impact in an adequate manner. 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 will use *subject normalised* observed and expected citation rates to avoid possible biases caused by subject-specific peculiarities or by different activity profiles. In particular, we suggest the application of the following three solutions for indicator selection and normalisation.

The first one is statistical-empirical and refers to the already mentioned computerised clustering of research institutes by their research profiles (cf. [THIJS & GLÄNZEL, 2008]). Some of the clusters represent specialised, others rather multi-disciplinary institutions with medical, technical or general profiles. We use these profile classes to analyse in how far field-normalisation is affected by specialisation and differentiation and to find an optimal field-depth for normalisation. The analysis is

based on 676 European universities and other research institutes with sufficiently large publication output.

The second one focuses on the citation windows used. In bibliometric practice, three-, four- or five-year citation windows are most commonly used. A second question arises concerning robustness and reliability of these subject-normalised and relative indicators at this level of aggregation. In order to answer this question, two different windows, a three-year and a five-year citation window, are applied to the above-mentioned "sample" of European institutes.

Third, although subject-normalised indicators to a large extent compensate for the above-mentioned biases they still remain measures of mean values. The question arises of how to capture and to visualise further important aspects of citation impact, for instance, the high-end of research by means of subject-normalised indicators. In order to identify highly cited publications and to build subject-normalised high-end indicators, we rely on results of earlier studies by the authors. The method of 'characteristic scores and scales' [GLÄNZEL & SCHUBERT, 1988] provides self-adjusting thresholds. We suggest the use of two different thresholds depending on the underlying publication output. A validity check for the appropriate citation window is done in this case, too. The citation-mean related charts are supplemented by relational charts reflecting the relative share of highly cited papers published by research institutes and research teams as well as the share of citations they receive with respect to the international reference standard. This actually forms the third of the above-mentioned solutions.

In addition to the methodological part, we finally provide a comparative sample analysis of the citation impact of selected 24 research universities representing 12 European countries. The results lead to the conclusion that instead of any linear ranking of universities, a more detailed, complex analysis is necessary to grasp and to reflect even this one, however important, aspect of performance among the manifold of university activities. However, this exercise also substantiates that properly normalised relative indicators measuring both mean citation impact and high-impact are suited to plot important aspects of research performance even across units with specialised, inter- and multi-disciplinary profiles.

Subject classification and profile classes

Data sources and data processing

All data used for this study were extracted from the yearly updates of the *Web of Science* database of Thomson-Scientific (Philadelphia, PA, USA). Only papers of the document types 'article', 'letter', 'note' and 'review' indexed in the 1999–2001 volumes have been selected. After a detailed cleaning, the bibliographic data have been processed to bibliometric indicators. Publications were assigned to countries and

institutions according to the address in the by-line of the paper. National thesauri produced at the Steunpunt O&O Indicatoren (SOOI) at K.U. Leuven on the basis of corporate address data have been used to identify and to assign publications to European institutions.

Subject classification of the publications was based on the field assignment of journals according to sixteen major fields in the sciences, social sciences and humanities developed in Leuven and Budapest [GLÄNZEL & SCHUBERT, 2003]. These fields are Agriculture & Environment, Biology (Organismic & Supraorganismic Level), Biosciences (General, Cellular & Subcellular Biology, Genetics), Biomedical Research, Clinical and Experimental Medicine I (General & Internal Medicine), Clinical and Experimental Medicine II (Non-Internal Medicine Specialties), Neuroscience & Behaviour, Chemistry, Physics, Geosciences & Space Sciences, Engineering, Mathematics, Social Sciences I (General, Regional & Community Issues), Social Sciences II (Economical & Political Issues) and Arts & Humanities. Although all domains were used for the determination of institutional publication profiles, the actual bibliometric analysis reported in this paper is restricted to the twelve fields in the sciences.

Subject diversification and profile analysis of research institutions

Unlike in the macro case, where the national research systems of medium-sized and big countries are principally expected to reflect multidisciplinarity, the question of diversification and specialisation becomes an important issue in the evaluation of institutional research. The breakdown by research fields for comparative analysis is, of course, possible at any level of aggregation but this approach alone does not capture and reflect the complexity and subject diversification of the institutions' research activities in an adequate manner. In order to overcome this deficiency, we have designed a solution for identifying institutions with similar research profiles. In earlier papers [LETA & AL., 2006; THIJS & GLÄNZEL, 2008], we have developed a method to classify and map the European and Brazilian institutional landscape on the basis of the individual institutes' research profiles. In what follows, this method, which consists of three steps, is briefly summarised to impart an understanding of the profile and data structure underlying the actual analysis.

The first step in our procedure was the breakdown of the publication output of each institute into research fields in order to construct their individual publication profiles. For this exercise, we have used the above-mentioned 16 major fields of the sciences, social sciences and humanities developed at KU Leuven and ISSRU–Budapest [GLÄNZEL & AL., 2003]. Each research profile can actually be considered a vector holding the share of each of these 16 fields in the total set of publications of the institute in question. Normalisation of data guarantees that the particular size of the total

publication output an institute produces within a certain time frame has no effect on the profile except for institutes with a very low publication activity as their share comes close to 0 or 1 for some fields. In order to avoid such distortions, small institutes with publication output below a given threshold have been removed (see [THIJS & GLÄNZEL, 2008]). In a second step, a cluster analysis using the Ward algorithm with squared Euclidean distances has been conducted on these research profiles. The Je(2)/Je(1) index introduced by DUDA & HART [1973] was used to find the optimum solutions. Beyond the almost trivial case of two clusters (medical and non-medical institutes) we have found eight clusters as second optimum solution. Finally, a predictive model was created using discriminant analysis. This model allows predicting group membership of research institutes based on their research profile as well as extending the classification to those institutes which were originally excluded because of their small publication output. The final classification is presented in Table 1. Although almost half the institutions belongs to a group with a rather specialised profile (see [THIJS & GLÄNZEL, 2008], the majority (about 80%) of all papers are published by institutions of the multidisciplinary clusters 3 and 5 (cf. [THIJS & GLÄNZEL, 2009]). This result did not strike us unexpectedly since diversification, that is, simultaneous activity in several science fields, normally goes with larger publication output than specialisation does. We would like to stress that all clusters presented in Table 1 still have a considerable overlap in their research profiles. This is a precondition for building relative citation indicators for possible application to both intra- and inter-class comparison. Research performance of institutions with (almost) completely different profiles should, however, not be subjected to direct comparisons since, for instance, comparing a medical university with a business school still remains an exercise of "comparing apples and oranges" and would therefore not make sense.

In the methodological section we will use these profile classes to analyse to what extent field-normalisation is affected by specialisation and differentiation. Therefore, we use the profile clusters only to answer the question of which field-depth should preferably be used for subject-based normalisation of citation indicators to meet the requirements of diversification and specialisation at the institutional level.

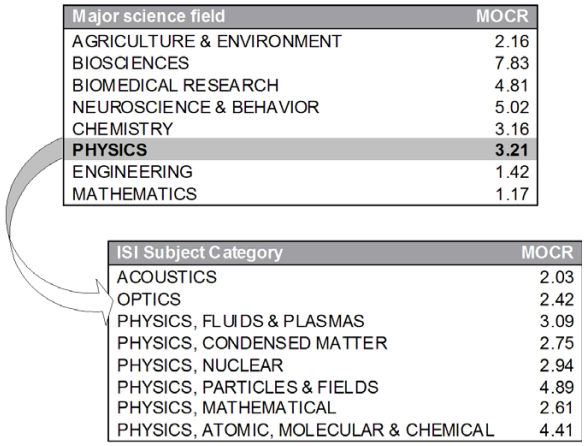
Table 1. The eight clusters resulting from the second optimum solution

Cluster	Code
Cluster 1 *(Biology)*	BIO
Cluster 2 *(Agriculture)*	AGR
Cluster 3 *(Multidisciplinary)*	MDS
Cluster 4 *(Geo & Space Science)*	GSS
Cluster 5 *(Technical & Natural Sciences)*	TNS
Cluster 6 *(Chemistry)*	CHE
Cluster 7 *(General & Research Medicine)*	GRM
Cluster 8 *(Specialised Medicine)*	SPM

Methods and results

Field-depth analysis for indicator normalisation

It is a commonly known fact that subject-specific peculiarities of publication and citation behaviour of scientists result in considerably differences in the standards of publication activity and citation-impact of the different subject areas. A glance at Table 1 already reveals that we have to expect different citation standards within the individual profile clusters as well. A further question arises from the classification shown in Table 1, in particular, whether research profiles of the individual fields possibly differ among clusters. In other words, has chemistry or physics research the same profile in the chemistry cluster #6 as in the multidisciplinary cluster #3 or the technical & natural sciences cluster #4? A similar question can be addressed regarding medical research in clusters #3, 7 and 8. Although this question seems to be rather rhetorical, in practice it might become quite significant if publication output of different clusters is broken down by fields and these fields serve as the basis of inter-cluster comparisons. The reason therefore lies in the heterogeneity of major fields in terms of scientists' communication behaviour. The field of physics might just serve as an example for this effect. Figure 1 first presents the mean citation rate of papers published in eight selected major fields in 2001 based on the 3-year citation window 2001–2003. The large differences among the citation impact of the various fields with biosciences (7.83) at the top and mathematics (1.17) at the bottom are apparent (see upper part of Figure 1). The lower part of Figure 1 shows the citation impact of eight selected subfields of physics for the same publication/citation period. The deviations of citation impact of physics subfields from each other are according to expectations less dramatic than those among the major fields shown in the upper section of the figure but they nevertheless visualise the heterogeneity of physics in terms of the authors' citation behaviour convincingly. The science fields and subfields that we have used in this example are taken from the hierarchical Leuven/Budapest subject classification scheme created by GLÄNZEL & SCHUBERT [2003] on the basis of journal assignment and using the ISI Subject Categories as lowest hierarchic level. The highest level consists of 12 major fields in the natural sciences, life sciences, applied sciences and mathematics (see in the 'Subject classification and profile classes' section). The level in between these major fields and the ISI Subject Categories comprises 60 subfields. Thus, a science field has five subfields on average and a subfield aggregates about three ISI Subject Categories each. The detailed hierarchical structure of fields and subfields can be found in the study by GLÄNZEL & SCHUBERT [2003].



Major science field	MOCR
AGRICULTURE & ENVIRONMENT	2.16
BIOSCIENCES	7.83
BIOMEDICAL RESEARCH	4.81
NEUROSCIENCE & BEHAVIOR	5.02
CHEMISTRY	3.16
PHYSICS	**3.21**
ENGINEERING	1.42
MATHEMATICS	1.17

ISI Subject Category	MOCR
ACOUSTICS	2.03
OPTICS	2.42
PHYSICS, FLUIDS & PLASMAS	3.09
PHYSICS, CONDENSED MATTER	2.75
PHYSICS, NUCLEAR	2.94
PHYSICS, PARTICLES & FIELDS	4.89
PHYSICS, MATHEMATICAL	2.61
PHYSICS, ATOMIC, MOLECULAR & CHEMICAL	4.41

Figure 1. Variation of mean citation impact among eight selected major fields and among subfields within a selected science area

In an earlier paper, THIJS & GLÄNZEL [2009] have illustrated using the example of chemistry, that the subfield profiles of chemistry research do indeed vary considerably among the individual clusters. Thus *physical chemistry* was, for instance, almost absent in the medical clusters whereas in the Technical- & Natural-Sciences and the Chemistry clusters *organic and medicinal chemistry* was less predominant than in the medical groups. The analysis of institutional research performance by subfields or even by ISI Subject Categories is less practicable at these lower aggregation levels because of the often small publication sets resulting from the breakdown. Although the creation of major profile clusters and/or the use of field-specific analysis might help avoid comparing 'apples and oranges', it is does by far not suffice to eliminate all profile-specific biases. The question arises therefore of how and at which level citation data are to be normalised in order to 'equalise' the above-mentioned citation biases. The idea of using different 'zoom' levels for cross-level comparisons was already addressed and described by ZITT & AL. [2005] and ADAMS & AL., [2008]. In the present study we proceed from a normalised citation indicator that has successfully been used since the 1980s, particularly, the subject-based *Normalised Mean Citation Rate* (NMCR, see, for instance, [BRAUN & GLÄNZEL, 1990]). This indicator is actually an extension of Vinkler's *Relative Subfield Citedness* index [VINKLER, 1986]. In this context we also mention that a similar measure (CPP/FCSm) is used at CWTS (cf. [MOED & AL., 1995]). These indicators have in common that the mean observed citation rate is gauged against its individual expectation on the basis of the subject to which it belongs. However, in the light of the above brief discourse on subject-related issues, the question arises of which hierarchic subject level should preferably be used as reference standard

for the definition of such an indicator to measure institutional citation impact. In Figures 2, 3 and 4, we plot and compute the relationship between the NMCR values at different subject-related aggregation levels for a sample of 676 European universities and research institutes.

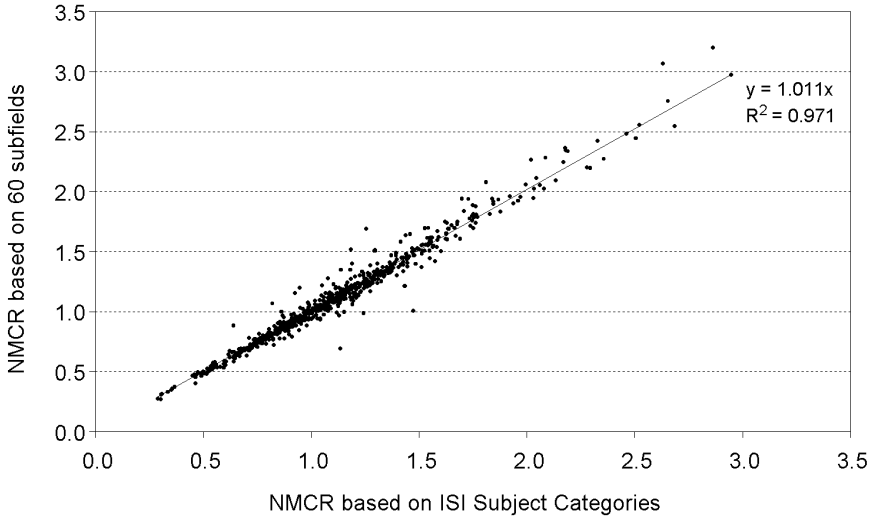


Figure 2. Plot of NMCR based on subfields vs. ISI Subject Categories for 676 European universities and research institutions

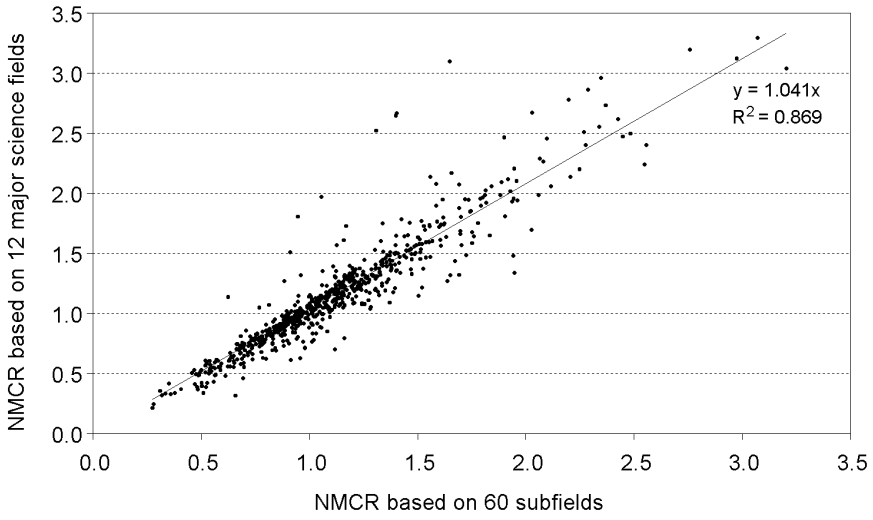


Figure 3. Plot of NMCR based on major fields vs. subfields for 676 European universities and research institutions

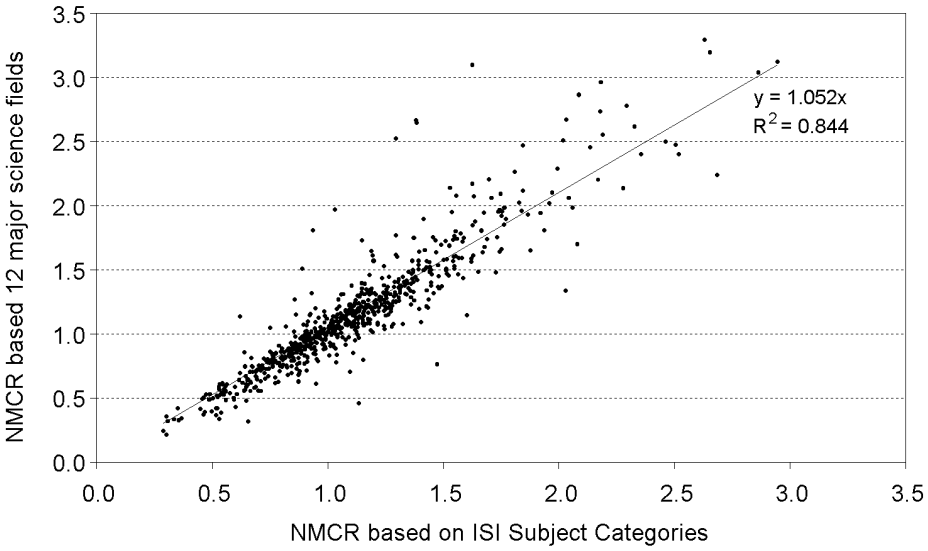


Figure 4. Plot of NMCR based on major fields vs. ISI Subject Categories for 676 European universities and research institutions

The examples suggest that the highest level, that is, the major subject fields might be too coarse for this exercise. On the other hand, one intuitively expects that using the lowest level, providing a fine-grained subject assignment, might be the most accurate approach. However, taking into account that many journals covered by the Web of Science are in practice assigned to 3, 4 or even more ISI Categories and that the assignment of publications to subfields is by far less fuzzy, it becomes conceivable that subfields could serve as the favoured reference level. A consequence of multiple assignments is the necessity of fractionating subject assignment and thus of calculating weighted averages for the corresponding individual field-expected citation rates. The weighting of fractional data is correct and fair if the sum of the individual field-expected citation rates over all publications in the system equals the citation total. Such averaging of subject standards still occurs at the level of subfields and major fields but to a much lesser extent than in the case of the ISI Categories. This issue therefore warrants close attention and scrutiny.

In the following paragraphs, we present a systematic comparison of the subject-normalised citation impact of European research institutions on the basis of the three different hierarchic levels of subject assignment. For this exercise we have selected those 676 organisations which have published at least 50 papers in the period 1999–2001 out of the total set of about 2000 European institutions. We have applied a

three-year citation windows beginning with the publication year, and thus shifted the windows by one year. We summed publication and citation counts to obtain the indicators for the periods 1999–2001 through 2001–2003 for the underlying publication years 1999, 2000 and 2001. Figures 2 through 4 present the pairwise linear regressions between the *Normalised Mean Citation Rate* of the 676 institutions calculated on the basis on the three different subject levels. We have set the intercept equal to 0 since NMCR becomes 0 if the underlying publication set is uncited, independent of the chosen reference level. Nonetheless, the application of an alternative regression model found the intercept for the three regressions at 0.028, 0.049 and 0.080, respectively. Our choice does therefore not result in any violation of reality.

Although the correlations between the different "NMCR models" are expectably strong since one might expect even identicalness of the NMCR values calculated in different ways (this identicalness would occur when each journal would uniquely belong to an ISI category), we have found the strongest correlation between the subfields and the ISI Categories. The other two correlations were somewhat weaker and, according to our expectations, the correlation between the lowest and the highest hierarchical level provided the poorest results. This result is to be expected and explained by the aggregation dynamics underlying the three hierarchical levels. Above all, the large number of outliers (cf. Figures 3 and 4) might cause problems in practice. Since the variables are not independent we have used the simple d measure already applied in similar context in earlier studies (e.g., [GLÄNZEL, 2000]) to measure the dispersion. d is defined as the variables' average deviation or distance. In the first case we have $d = 0.042$, whereas d exceeds 0.100 in the two other cases (0.104 and 0.117, respectively).

The results reported in Figures 2–4 are based on all institutions with at least 50 publications each, irrespective of their belonging to any particular profile cluster. The patterns presented in Figures 2–4 become, however, even more distinct if the eight profile clusters are studied separately. While the linear regression model provides almost the same patterns for the multidisciplinary cluster (#3), we have found Geo & Space Science (#4) the most problematic one. We present the scatter plots and the linear regression for both clusters as well as for cluster #1 (Biology) in Figure 5. Only institutions with at least 50 papers in the period 1999–2001 have been taken into consideration here, too. While the correlation between the subfields and the ISI Categories is still acceptable in the case of cluster #4, the NMCR calculated on the basis of major fields and the ISI Categories for this cluster results in rather weakly correlated variables. The fact that the slope of 1.4 distinctly exceeds the value of 1.0, clearly demonstrates the inconsistence of the major-field model (see Figure 5, bottom right).

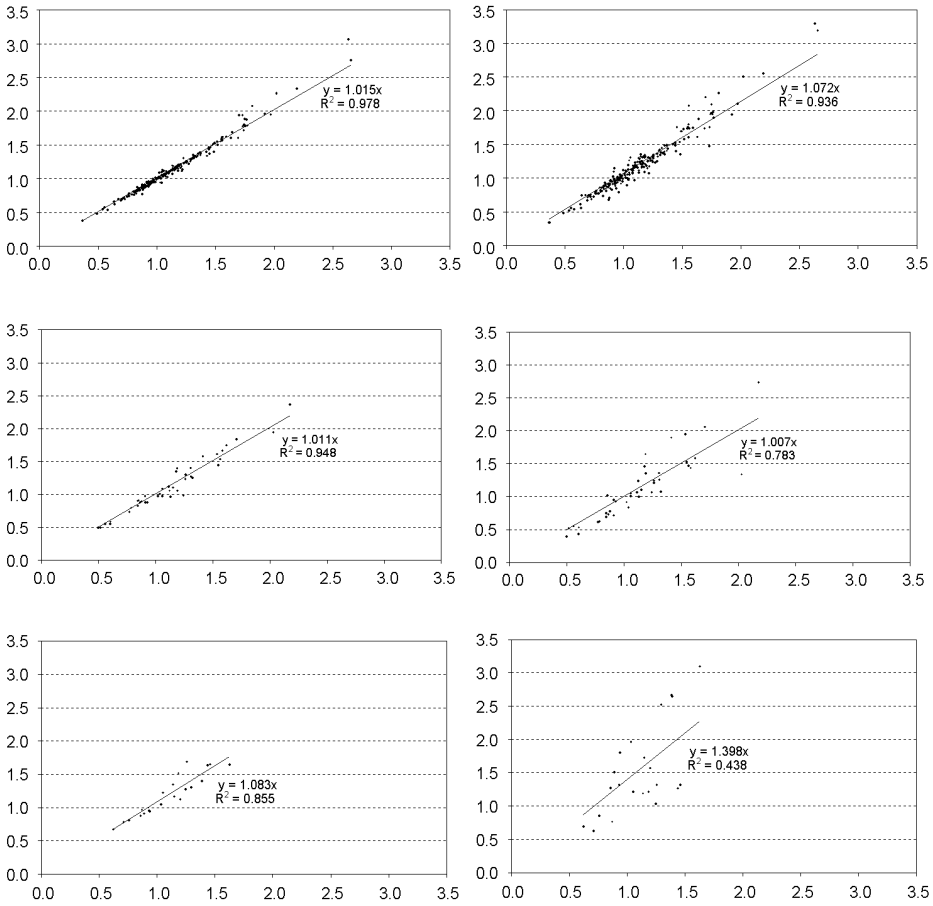


Figure 5. Plot of NMCR based on subfields [y] vs. ISI Subject Categories [x] (left) and major fields [y] vs. ISI Subject Categories [x] (right) for Cluster 3 (top), Cluster 1 (centre) and Cluster 4 (bottom)

In addition to the above regression analysis, we have calculated the rank correlation using the Spearman coefficient for the largest 20 institutes in each cluster. The results are presented in Table 2. The correlation is generally strong, but normalisation based on subfields regularly provides 'better' results than the major-field based version. The smallest cluster #6 (chemistry) forms a certain exception to the rule; here the correlation is the weakest among all the clusters studied.

Summarising, we can conclude that we obtain consistent results for the calculation of subject-normalised citation indicators for both the ISI Subject Categories and the subfields. However, to avoid too much fractionation, weighting and averaging to

account for multiple subject assignments, we prefer the use of subfield-based standards. All further analysis will therefore be based on this approach.

Table 2. Spearman rank correlation ρ for the 20 largest institutions of each cluster

Cluster	60 vs. ISI	12 vs. ISI
1	0.961	0.914
2	0.901	0.859
3	0.995	0.976
4	0.901	0.659
5	0.991	0.941
6	0.839	0.659
7	0.943	0.917
8	0.967	0.931

Citation windows for the analysis of institutional research performance

Before we extend our considerations to the high-impact analysis, we clarify an important validity issue. In particular, the issue of which citation window should underlie the citation analysis becomes more significant, but also more critical, at the lower levels of aggregation. No doubt, a longer citation window results in a higher reliability of the indicators than a shorter one. Also, the often-heard argument that the usual windows of 3 to 5 year are not sufficient in specific fields of the applied sciences and in mathematics, suggests the application of as large as possible citation windows. On the other hand, science policy and research management is interested in the evaluation of the most recent research results. Bibliometric studies of the situation years ago are at best suited for providing background information or as an input to longitudinal or historical studies. The application of a five-year citation window already refers to research done seven or eight years ago. In particular, various time related considerations come into play. When selecting the relevant citation window, one indeed has to add the time necessary to conduct the research, the time to organise and to condense the results obtained into written documents, the time for the reviewing process, a certain publication delay dependent on the journal where the paper is published as well as the time for indexing the most recent citing literature in the citation index, and finally the time for processing all the necessary bibliographic information. Finally, at the level of research groups this might become critical if one takes into account that the constitution of a research team often considerably changes over a period of 6–8 years.

A second important issue relates to the ageing of the literature. In a study on the possibility and the reliability of predictions of citation processes [GLÄNZEL, 1997], it was shown that true predictions of future citation rates are indeed possible, if the initial

reference period is close to the prediction period. The goodness-of-predictions increases with the length of the initial period and decreases with the length of the interval to be predicted. The results of the above-mentioned study suggested the use of a three-year citation window as a good compromise between the fast reception of life science and technology literature and that of the slowly ageing theoretical and mathematical subjects. Furthermore, the choice of three years still allows the evaluation of recent research results, but is usually long enough to determine future citation impact.

The 3-year citation window has – besides longer windows – long successfully been used at different levels of aggregation (see, for instance, [MOED, 1996; GLÄNZEL, 2000; GLÄNZEL & SCHUBERT, 2003; SCHUBERT & GLÄNZEL, 2007; VAN RAAN, 2006A, 2006B]. Nevertheless, we should have a closer look at how sensitive subject-normalised citation indicators (based on the subfield level) are to different initial citation periods. We have used a three- and five-year window for this exercise. Figure 6 presents the scatter plot of the Normalised Mean Citation Rate based on a 5-year citation window vs. a 3-year window for the 676 European universities and research institutions with at least 50 papers each in the period 1999–2001. The practically perfect fit of the linear regression model with very strong correlation shows that at this level, too, a three-year window suffices to measure citation impact in an adequate manner. In qualitative terms, the growth of the observed citation rates on average parallels that of their subfield-based expectations. We will therefore use a three-year citation window in the following sections of this study.

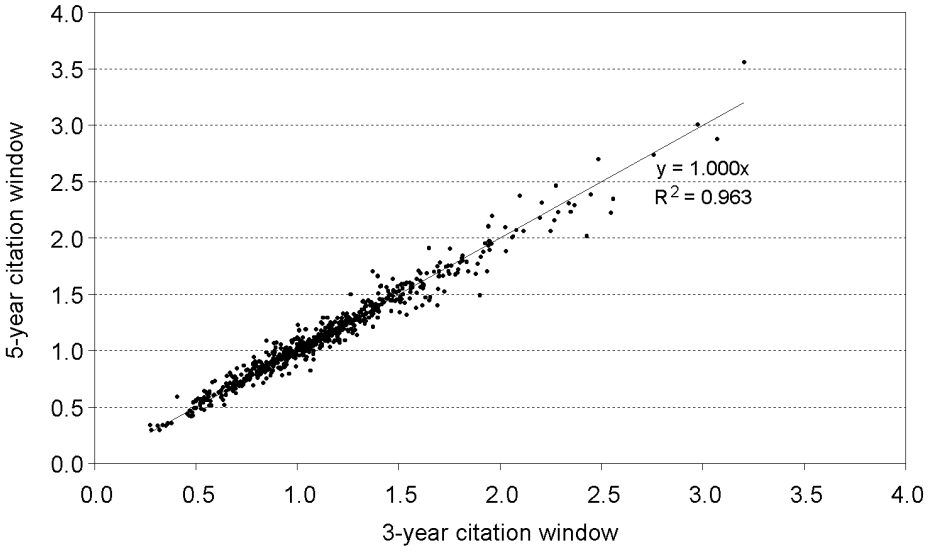


Figure 6. Plot of NMCR based on 5-year citation window vs. 3-year window for 676 European universities and research institutions

Characteristic scores and scales for highly cited publications

In the previous sections we have found a solution for measures of citation impact based on mean observations and expectations that are practically subject-insensitive and time-invariant. The question arises of how to capture and to visualise the high-end of research spectrum in a similar way of relative subject-independence and time-invariance. In order to answer this question, we first need a closer look at the main characteristics of citation distributions.

The power-law property of the tail of citation distributions can be used to derive a straightforward rule for the determination of highly cited articles from the method of characteristic score and scales [SCHUBERT & GLÄNZEL, 1988; GLÄNZEL, 2007]. The effect of cumulative advantage, which is typical of bibliometric processes, can actually be described with by Paretian distributions (e.g., [PRICE, 1976; TAGUE, 1981]). We will use this property to derive a straightforward rule for the determination of highly cited articles from the method of *characteristic score and scales* [GLÄNZEL & SCHUBERT, 1988; GLÄNZEL, 2007]. This method can be summarised as iteratively truncating samples at their mean value and recalculating the mean of the truncated sample until the procedure is stopped or the sample is empty. We outline the procedure by the example of citation rates of a given paper set preferably published on the same topic or in the same subfield. The conditional means calculated for the truncated samples are called *characteristic scores*, and are denoted by b_k, $k \geq 0$. We start with b_0, which equals 0 by definition, and b_1 being the mean citation rate of the papers in the given set. Now we discard all papers cited less than the mean. The mean citation rate of the remaining papers is denoted by b_2. Again, those papers cited less than b_2 are removed. The mean citation rate of the rest is b_3. This procedure is repeated till the set is empty or it is stopped at a given value k (see [GLÄNZEL & SCHUBERT, 1988]). We can define the following zones or classes whereby we usually stop at $k = 3$ or even earlier. Above that value, the low number of papers in the truncated sample might not allow a reliable citation analysis of the remaining set. Thus the procedure results in subdividing the original distribution into the following different classes or zones: $[0, b_1)$ is the class of 'poorly cited' papers, $[b_1, b_2)$ contains 'fairly cited' papers, $[b_2, b_3)$ and $[b_3, \infty)$ are the two classes of highly cited papers called 'remarkably cited' for $k = 2$ and 'outstandingly cited' papers for $k = 3$, respectively. Figure 7 visualises the selection procedure and the creation of the characteristic citation classes.

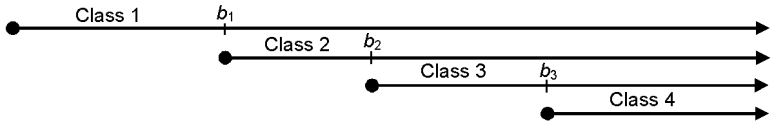


Figure 7. Visualisation of characteristic scores and scales for four classes

According to the characterisation theorem for Paretian distributions by GLÄNZEL & AL. [1984], the conditional means b_k can be approximated by $b_k \approx \{\alpha/(\alpha-1)\}\cdot b_{k-1} + b_1$, where α is the parameter of the underlying Pareto distribution. According to GLÄNZEL & SCHUBERT [1988], we obtain the approximation

$$b_k = \sum_{i=0}^{k-1} a^i \cdot b_1 = (a^0 + a^1 + \ldots + a^{k-1}) \cdot b_1$$

by recursion, where $k > 0$ and $a = \alpha/(\alpha-1)$. Empirical studies have shown that $\alpha \approx 2$ for citation windows of 3–5 years [SCHUBERT & GLÄNZEL, 2007; GLÄNZEL, 2008]. In this case we obtain $a = 2$ and thus $b_2 = 3b_1$ and $b_3 = 7b_1$ for the thresholds $k = 2$ and $k = 3$, respectively. In the present study, these thresholds are applied to the same 60 subfields as above. In particular, the citation impact of each individual paper is compared with the three-fold or seven-fold of the corresponding subject standard, according as $k = 2$ or $k = 3$ was chosen. The citation scores found this way have the following important properties.

As was shown in an earlier study [GLÄNZEL, 2007], the scores b_k are time-dependent and, of course, strongly sensitive to the subject matter as well but the share of papers in the individual zones is relatively stable over the citation window, and does not noticeably vary among different subfields either.

Since the k thresholds are applied to the same subfields as the above expected and observed citation rates, indicators on highly cited papers defined on the basis of characteristic scores and scales can as such be considered *subfield normalised*, and therefore be used as direct supplement to the previous set of relative indicators.

According to our observations, the reference standard of the share of outstandingly cited papers varies between 1% and 2% and that of remarkably cited papers amounts to about 5% (e.g., [GLÄNZEL, 2007]). As a consequence of this rule, we have to restrict studies of highly cited papers to institutions with at least 1000 papers each (outstandingly cited papers) or, alternatively, to those with no less than 200–250 papers in the case of remarkably cited papers. Otherwise, the institutional sets of highly cited papers are not expected to have more than 10 publications each. Paper sets of such small size are, however, no longer appropriate for the statistical citation analysis attempted in this study.

The following regression analysis substantiates the above mentioned stability properties on the sample of 132 European universities and other research institutions (derived from the start sample of 676 institutions) with individual publication output of at least 1000 papers each. We have calculated the share of outstandingly cited papers in the institutional total (i.e., $k = 3$) for papers published in the period 1999–2001 based on a 3-year citation window and a 5-year citation window beginning with the publication year for each paper. Figure 8 presents the results of the linear regression model. The correlation is strong and the slope (0.959) is quite close to the expected value of 1.0 as well. From the

example of medium-sized and large institutions, we can conclude that the 3-year citation window is sufficiently long for the analysis of highly cited papers, too.

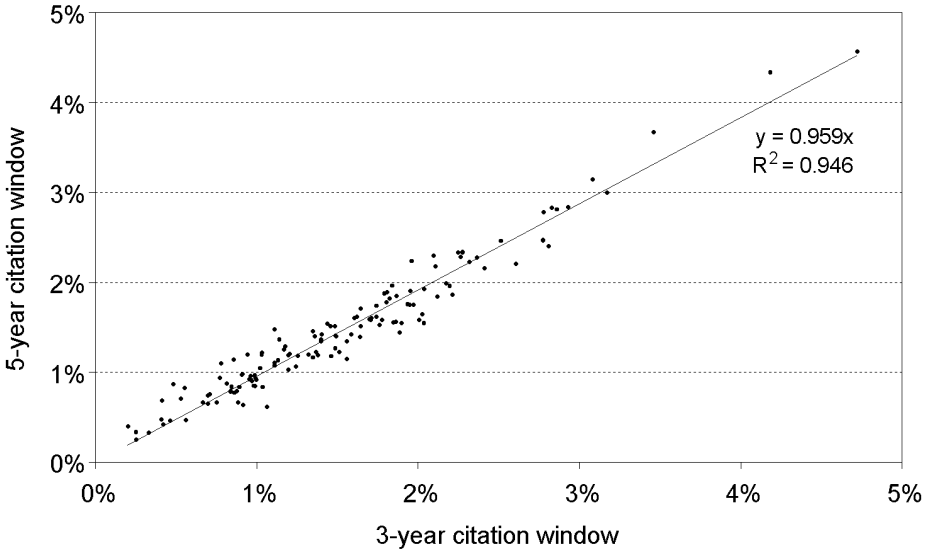


Figure 8. Share of highly cited papers in all papers of 132 institutions with at least 1000 publications each based on two different citation windows

Subfield-normalised relational charts for comparative assessment of citation impact

In this section we finally apply the above methodological results to a specific selection of research universities. For this exercise, we have selected two large universities per country from twelve medium-sized European countries. In particular, the following countries have been chosen: Austria, Belgium, Denmark, Finland, Hungary, Ireland, Italy, Netherlands, Portugal, Spain, Sweden and Switzerland. Since the largest universities (in terms of publication output) usually have multidisciplinary research profiles, we replaced some of the originally selected universities by medical or technical universities with a similarly large publication output. As a consequence, we ended up having two medical universities and three technical universities in the selection.

The publication output of the selected institutions ranges between about 1200 and 8500 papers in the period 1999–2001. Since the selection is based on the above-mentioned 'correction,' and hence might be considered somewhat arbitrary, and the

intention of this study is to merely present the methodology using the institutions as samples, the universities will be represented anonymously by letters. The two medical universities are U and X, the technical universities are G, T and Y. A comparative study of a more exhaustive and non-anonymous selection of universities is in preparation. Before we present the citation data, we briefly summarise all the indicators used within the present comparative assessment. All indicators are based on three-year citation windows.

Observed citation rates

Mean Observed Citation Rate (MOCR). MOCR is defined as the ratio of citation count to publication count. It reflects the factual citation impact of the unit under study, which might be a country, region, institution, research group etc. If the unit under study has n papers in the given publication period and the ith paper (i = 1, 2, … n) has attracted c_i citations, then one can write

$$MOCR = \frac{\sum_{i=1}^{n} c_i}{n}.$$

Expected citation rates

Mean Expected Citation Rate (MECR). The expected citation rate of a single paper is defined as the average citation rate of all papers published in the same journal in the same year. Instead of the one-year citation window to publications of the two preceding years as used in the Journal Citation Report (JCR), any other citation window can be used, as explained above. However, the same combination of publication period and citation window as above has to be applied for comparison with the corresponding observed citation impact. Using the above notation and assuming that the impact measure of the journal where the ith paper of the unit appeared is x_i, one has

$$MECR = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Field Expected Citation Rate (FECR). Analogously to the previous indicator, the field-expected citation rate of a single paper is defined as the average citation rate of all papers published in the same subject in the same year. In order to obtain valid results, the same publication period and citation window has to be used as in the case of the previous to indicators. Analogously to the previous case one has

$$FECR = \frac{\sum_{i=1}^{n} f_i}{n},$$

where f_i is the weighted average of the impact of those subfields to which the ith paper was assigned.

Relative citation rates

The ratio of the two previous indicators (MECR/FECR) expresses whether the unit under study publishes on average in higher (lower) impact journals than expected on the basis of the subfields where the unit was active.

*Normalised Mean Citation Rat*e (NMCR). This indicator is defined as the ratio of the observed and field-based expected citation impact, that is,

$$NMCR = \frac{MOCR}{FECR} = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} f_i}.$$

NMCR gauges citation rates of the papers against the standards set by the specific subfields. Its neutral value is 1 and NMCR >(<) 1 indicates higher(lower)-than-average citation rate than expected on the basis of the average citation rate of the subfield. Due to its definition, this measure is largely insensitive to the differences between the citation practices of the different science fields and subfields. NMCR has been introduced by BRAUN & GLÄNZEL [1990] in the context of national publication strategy. A similar measure (CPP/FCSm) is used at CWTS (cf. [MOED & AL, 1995]).

Relative Citation Rate (RCR). RCR is defined as the ratio of the observed and journal-based expected citation impact, that is,

$$RCR = \frac{MOCR}{MECR} = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} x_i}.$$

This indicator measures whether the publications of the unit under study attract more or less citations than expected on the basis of the journal impact measures, i.e., the average citation rates of the journals in which they appeared. Since the citation rates of the papers are gauged against the standards set by the specific journals, also this indicator is insensitive to the different citation practices in the various science fields and subfields. Analogously to the NMCR, RCR = 0 corresponds to un-citedness, RCR < 1

means lower-than-average, RCR > 1 higher-than-average citation rate, RCR = 1 if the set of papers in question attracts just the number of citations expected on the basis of the average citation rate of the publishing journals. RCR was the preferred relative indicator at ISSRU (Budapest) and has often been applied to comparative macro and meso studies ever since (see, e.g., [BRAUN & AL., 1985; SCHUBERT & BRAUN, 1986; SCHUBERT & AL., 1989]. It should be mentioned that a version of this relative measure, namely, CPP/JCSm is used at CWTS in Leiden (see [MOED & AL., 1995].

High-Impact Activity is the ratio of the unit's share of highly cited papers in all papers and the corresponding world standard. This measure has been introduced by GLÄNZEL & SCHUBERT [1992], but has in the past been applied to a different definition of high impact. In this study, we base the measure of high citedness on the definition given in the previous section, and actually use the threshold of $k = 3$ to select outstandingly cited papers for the high-impact analysis.

High-Impact Attractivity is defined analogous to the previous indicator, particularly, as the ratio of the unit's share of citations attracted by its highly cited papers in all citations received by the unit under study and the corresponding world standard. The same conditions and thresholds as in the definition of the previous indicator have to be used.

Non-negative relative indicators are preferably presented in a *relational chart*, that is, in the right-upper quadrant of a simple two-dimensional Cartesian coordinate system (see [SCHUBERT & BRAUN, 1986]). The relative indicator itself is obtained as the slope of the straight line connecting the origin (0, 0) with the corresponding point (x, y), where x represents the expectation and y the observation. Thus the line $y = x$ indicates the balance between observation and expectation. In the new relational charts designed for the presentation of normalised indicators, the upper-right quadrant is further subdivided into six sections by adding to the bisector the two straight lines $x = 1$ and $y = 1$ indicating two other equilibria, in particular, the conformity with the corresponding underlying reference standards [BRAUN & GLÄNZEL, 1990]. A prototype of those relational charts as well as the interpretation of the six sectors is presented in Figure 9. Based on the validation analyses and on the normalised indicator base described above, we propose the use of three charts to support a comparative institutional assessment of citation impact.

In a first step, the scatter plot of (MECR, MOCR) values is presented in a traditional relational chart (see Figure 10). Both expectation and observation cover quite a large range of citation impact. The fact that U is a medical university whereas Y 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 U considerably exceed those of the technical university G. We observe a similar situation for X and Y, however, on a much lower level. Also T, as a technical university, appears in the low-end group of this diagram.

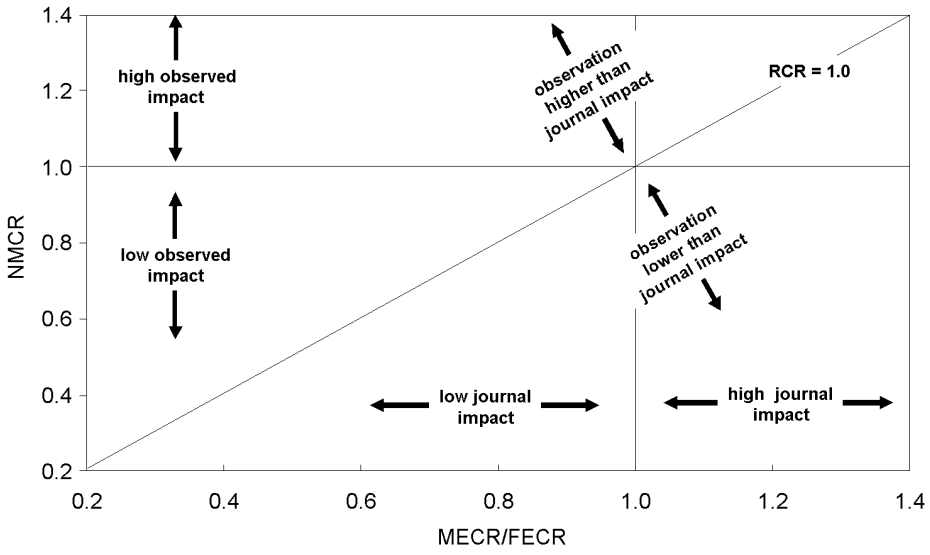


Figure 9. Prototype of a subject-normalised relational chart plotting NMCR vs. MECR/FECR

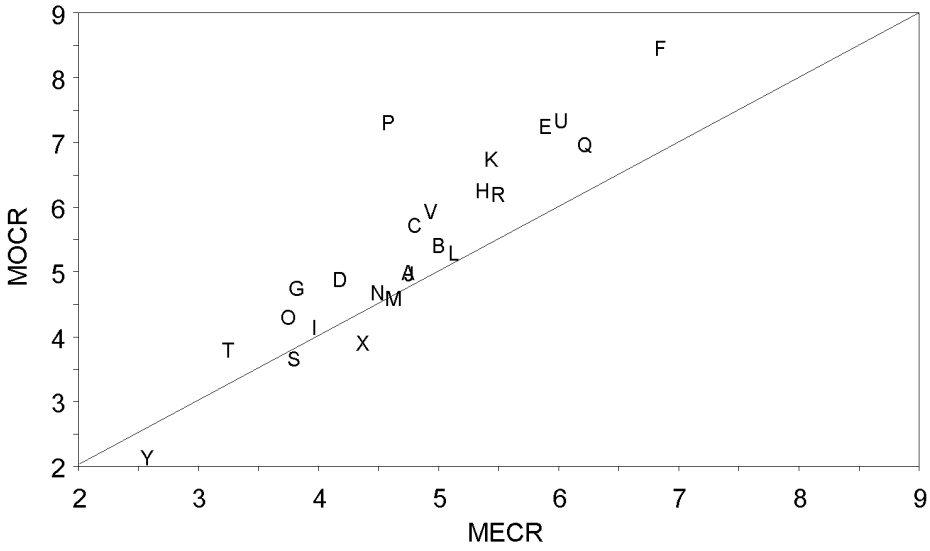


Figure 10. Relational chart of expected and observed citation rate for 24 selected universities

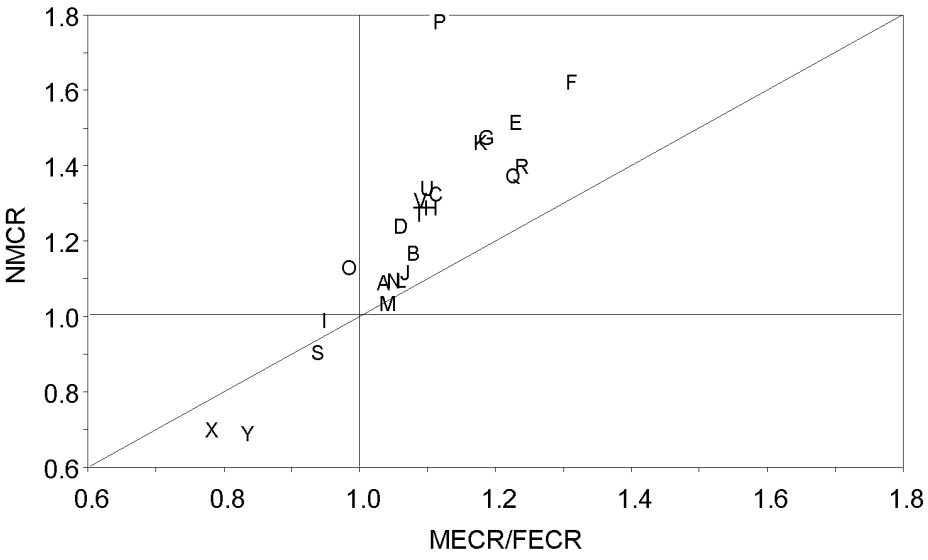


Figure 11. Subfield-based relational chart for 24 selected universities

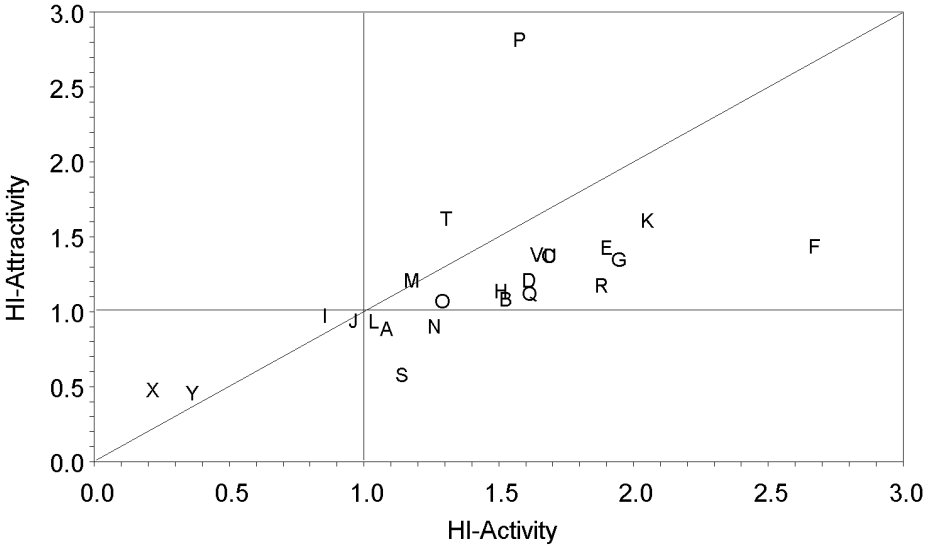


Figure 12. Relational chart of High-Impact Activity and Attractivity for 24 selected universities (the diagonal line indicates equality of High-Impact Activity and Attractivity, the horizontal and vertical lines indicate conformity between observation and the corresponding expectation)

We obtain a completely different result and insight if subfield-based normalisation is applied. The plot of subfield-normalised observation against subfield-normalised expectation is presented in Figure 11. The positions of universities G and U have interchanged; the same applies to X and Y. The effect of field-specific lower impact of technical universities and the usual high impact of medical of medical universities has thus been eliminated. X and Y take almost the same position in the lower-left section of the chart. The position of university T has changed as well; in the new chart it has joined the group in the most advantageous section. Some changes also affect universities P and F, both, however, with outstanding citation impact.

In order to deepen these results we present the plot of high-impact activity vs. high-impact attractivity for these universities (see Figure 12). This chart mirrors the previous diagram apart from that the attractivity of the high-end group does not quite meet the expectation set by their high-impact activity. Though, as well the less favourable situation of universities X and Y as the outstanding citation impact of F and especially that of P are confirmed by this diagram.

The relational charts, confronting non-normalised and normalised citation impact assessments, therefore offer interesting insights and findings, moving away from the simplified linear ranking and evaluation methods that are often advocated and used today. They complement one another in that they allow for a much more fine-grained assessment of citation impact, putting every institution to be evaluated in perspective to those institutions that share a similar disciplinary profile.

Conclusions

Both the three-level hierarchical subject-classification scheme based on the ISI Subject Category system and the institutional profile clusters developed at SOOI/KULeuven have been used to determine the appropriate field-depth for the calculation of subject-normalised citation indicators. The level of 60 subfields proved to be an acceptable choice and was able to provide stable and consistent results. The normalised indicators can be used for both intra- and inter-cluster comparative analysis as well as for domain studies at this level of aggregation. In addition to the relative citation indicators, the high-end of institutional research is best reflected by highly cited papers. The thresholds can readily be derived from the method of 'characteristic score and scales'. Two thresholds are actually suggested, depending on the underlying publication output. The study has also shown that the use of a three-year citation window suffices for building both relative and high-impact citation indicators.

Although this study aimed, above all, at solving problems arising from the evaluation of institutional research performance, it should be stressed that the methodological approach described in the previous section is, without any restriction, appropriate for use at the macro level as well. At a lower level of aggregation, some

limitations concerning the application to the comparative analysis of research groups can arise from the possibly missing "critical mass" of highly cited papers and the sometimes diverging research profiles of different groups. Statistical tests for the significance of the deviation from the corresponding expectations (see [SCHUBERT & GLÄNZEL, 1983]) can, on the other hand, serve as an additional option at this level where the standard errors of mean can – due to the relatively small publication sets – become quite considerable. Finally, both the method developed in this paper and its application to 24 universities highlight the need for well-thought out and careful approaches to benchmark research institutions on the basis of their citation impact. The relational charts designed and discussed in the paper illustrate how insights obtained from citation impact mapping change according to the underlying indicator base being used. Normalised indicators are highly necessary if one does not want to compare "apples" and "oranges". Comparative assessments therefore need sufficient levels of indicator sophistication. We are convinced that the normalised indicator base and its potential for relational mapping, as developed in this paper, offer this highly needed level of accuracy and sophistication.

References

ADAMS, J., K. GURNEY, L. JACKSON (2008). Calibrating the zoom – a test of Zitt's hypothesis, *Scientometrics*, 75 (1) : 81–95.

BRAUN, T., W. GLÄNZEL, A. SCHUBERT (1985), *Scientometric Indicators. A 32-Country Comparison of Publication Productivity and Citation Impact. World Scientific*, Singapore – Philadelphia.

BRAUN, T., W. GLÄNZEL (1990), United Germany: The new scientific superpower? *Scientometrics*, 19 (5–6) : 513–521.

DUDA, R. O., P. E. HART (1973), *Pattern Classification and Scene Analysis*. New York: Wiley.

GLÄNZEL, W., A. TELCS, A. SCHUBERT (1984), Characterization by truncated moments and its application to Pearson-type distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66 : 173–183. (Correction: *Probability Theory and Related Fields*, 74 (1987) 317.)

GLÄNZEL, W., A. SCHUBERT (1988), Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, 14 : 123–127.

GLÄNZEL, W., A. SCHUBERT (1992), Some facts and figures on highly cited papers in the sciences, 1981–1985, *Scientometrics*, 25 (3) : 373–380.

GLÄNZEL, W. (1997), On the reliability of predictions based on stochastic citation processes, *Scientometrics*, 40 (3) : 481–492.

GLÄNZEL, W. (2000), Science in Scandinavia: A bibliometric approach, *Scientometrics*, 48 (2) : 121–150. (Correction: *Scientometrics*, 49 (2) (2000) 357)

GLÄNZEL, W., A. SCHUBERT (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes, *Scientometrics*, 56 (3) : 357–367.

GLÄNZEL, W. (2007), Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation, *Journal of Informetrics*, 1 (1) : 92–102.

GLÄNZEL, W. (2008), On some new bibliometric applications of statistics related to the h-index, *Scientometrics*, 76 (3) forthcoming

LETA, J., W. GLÄNZEL, B. THIJS (2006), Science in Brazil. Part 2: Sectoral and institutional research profiles, *Scientometrics*, 67 (1) : 87–105.

MOED, H. F., R. E. DE BRUIN, TH. N. VAN LEEUWEN (1995), New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications, *Scientometrics*, 33 (3) : 381–422.

MOED, H. F. (1996), Differences in the construction of SCI based bibliometric indicators among various producers: A first over view, *Scientometrics*, 35 (2) : 177–191

PRICE, D. J. DE SOLLA (1976), A general theory of bibliometric and other cumulative advantage processes, *Journal of the American Society for Information Science*, 27 (5–6) : 292–306.

REIST-2 (1997), *The European Report on Science and Technology Indicators 1997*. EUR 17639. European Commission, Brussels.

SCHUBERT, A., W. GLÄNZEL (1983), Statistical reliability of comparisons based on the citation impact of scientific publications, *Scientometrics*, 5 (1) : 59–74.

SCHUBERT, A., T. BRAUN (1986), Relative indicators and relational charts for comparative-assessment of publication output and citation impact, *Scientometrics*, 9 (5–6) : 281–291.

SCHUBERT, A., W. GLÄNZEL, T. BRAUN (1989), Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major fields and subfields 1981–1985. *Scientometrics*, 16 (1–6) : 3–478.

SCHUBERT, A., W. GLÄNZEL (2007), A systematic analysis of Hirsch-type indices for journals, *Journal of Informetrics*, 1 (3) : 179–184.

TAGUE, J. M. (1981), The success-breeds-success phenomenon and bibliometric processes, *Journal of the American Society for Information Science*, 32 (4) : 280–286.

THIJS, B., W. GLÄNZEL (2008), A structural analysis of publication profiles for the classification of European research institutes, *Scientometrics*, 74 (2) : 223–236.

THIJS, B., W. GLÄNZEL (2009), A structural analysis of benchmarks on different bibliometric indicators for European research institutes based on their research profile, *Scientometrics*, forthcoming.

VAN RAAN, A. F. J. (2006A) Statistical properties of bibliometric indicators: Research group indicator distributions and correlations, *Journal of the American Society for Information Science and Technology*, 57 (3) : 408–430.

VAN RAAN, A. F. J. (2006B), Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups, *Scientometrics*, 67 (3) : 491–502.

VINKLER. P. (1986), Evaluation of some methods for the relative assessment of scientific publications, *Scientometrics,* 10 (3–4) : 157–177.

ZITT, M., S. RAMANANA-RAHARY, E. BASSECOULARD (2005), Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation, *Scientometrics*, 63 (2) : 373–401.