

MICHEL ZITT^{1,2}, SUZY RAMANANA-RAHARY², ELISE BASSECOULARD¹

¹*Lereco, INRA, Nantes (France)*

²*Observatoire des Sciences et des Techniques (OST), Paris (France)*

zitt@nantes.inra.fr, ramanana@obs-ost.fr, bassecou@nantes.inra.fr

Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation

Summary

As citation practices strongly depend on fields, field normalisation is recognised as necessary for fair comparison of figures in bibliometrics and evaluation studies. However fields may be defined at various levels, from small research areas to broad academic disciplines, and thus normalisation values are expected to vary. The aim of this project was to test the stability of citation ratings of articles as the level of observation --- hence the basis of normalisation --- changes. A conventional classification of science based on ISI subject categories and their aggregates at various scales was used, namely at five levels: all science, large academic discipline, sub-discipline, speciality and journal. Among various normalisation methods, we selected a simple ranking method (quantiles), based on the citation score of the article in each particular aggregate (journal, speciality, etc.) it belonged to at each level. The study was conducted on articles in the full SCI range, for publication year 1998 with a four-year citation window. Stability is measured in three ways: overall comparison of article rankings; individual trajectory of articles; survival of the top-cited class across levels. Overall rank correlations on the observed empirical structure are benchmarked against two fictitious sets that keep the same embedded structure of articles but reassign citation scores either in a totally ordered or in a totally random distribution. These sets act respectively as a 'worst case' and 'best case' for the stability of citation ratings. The results show that: (a) the average citation rankings of articles substantially change with the level of observation (b) observation at the journal level is very particular, and the results differ greatly in all test circumstances from all the other levels of observation (c) the lack of cross-scale stability is confirmed when looking at the distribution of individual trajectories of articles across the levels; (d) when considering the top-cited fractions, a standard measure of excellence, it is found that the contents of the 'top-cited' set is completely dependent on the level of observation. The instability of impact measures should not be interpreted in terms of lack of robustness but rather as the co-existence of various perspectives each having their own form of legitimacy. A follow-up study will focus on the micro levels of observation and will be based on a structure built around bibliometric groupings rather than conventional groupings based on ISI subject categories.

Introduction

Citations and impacts have been recognised for decades as a feature of central importance in science studies. They have received even more attention recently with the pervasive practices of institutional assessment, benchmarking, and 'excellence' measurement. Interpretations of citation analyses are subject to many caveats which have been studied by both sociologists and bibliometricians from a variety of schools. A major issue is the discrepancy of citation behaviour across fields (Pinski & Narin, 1976, Murugesan & Moravcsik, 1978). In the early eighties various proposals for field-normalisation of impact figures were suggested, in both the USA and Europe, making comparisons possible between say, articles

in mathematics (a generally low-impact field) and in fundamental biology (a generally high impact field). Some milestones in bibliometric research were reviewed by Schubert & Braun (1996).

There is little doubt about the need of normalisation, but the question arises of the particular level that should be used. A narrow research area? A too small reference set can be statistically fragile and unstable over time. A large academic discipline? It may be too heterogeneous, hence inefficient for normalisation. Thus, various pros and cons of narrow versus large reference sets can be discussed (see the conclusion). To a certain extent, this corresponds to different perspectives having their own form of legitimacy. If we want to address the problem in general, we must consider a wide range of extensions of the reference set used for normalisation. In other words, we have to examine the sensitivity of normalised impact measures for particular articles as the scale of observation / normalisation changes. To this end, we need two pieces of information, first the citation score of individual articles (available in SCI series), and also a complete (i.e. multi-level) and realistic classification of scientific articles, which will provide, at various levels of aggregation, the reference set for normalisation or relative ranking.

There is no 'objective' way to uncover the structure of science, which may reflect institutional habits, mental representations or self-organisation phenomena. Among the possible ways of offering manageable classifications, there are three classical approaches. Firstly, the projection of institutional settings, for example traditional academic disciplines definition; secondly, the information retrieval categories in databases often based on experts' advice; thirdly, the clusters uncovered by bibliometric analyses of scientific networks (lexical and citation networks), with many sub-options, e.g. for citation networks: citation transactions, co-citations, bibliographic coupling. These broad families of methods are likely to provide different views of the structure of science.

In the present study we rely on the ISI retrieval categories and we will discuss the issues from the macro-perspective alone. Macro-analysis is defined here as starting from the journal level and to build the further levels as collections of journals, whereas we define micro-analysis by the use of clusters built from the document level, regardless of journals. The best-known example of macro-classification is ISI's 'subject category' list. Subject categories are far from perfect, based on rather unclear delineation methods, but since they are widely used and their journal lists easily available, they are often used as proxies for specialised research areas. We will use these ISI categories as the next level of aggregation, the first being the journal. Then we will gather these specialities into 'sub-disciplines' and finally sub-disciplines into disciplines. Including the all-science level, we have five levels of aggregation that will be used as normalisation referents. Certain limitations of this macro perspective, based on ISI categories and their further aggregations into our own schemes, are apparent. The hierarchical structure is meant to be realistic, but it cannot escape some arbitrary choices. We will come back to this subject in the conclusion.

We have chosen to study the stability by relative rankings (quantiles), instead of cardinal normalised impacts, at all levels of aggregation. In addition, we focus on a particular class, the 'highly-cited' articles. Among the many approaches to measure excellence, measures relying on highly-cited papers can be the most tempting, this in spite of the many difficulties of interpretation raised by both sociologists and bibliometricians. 'Excellence' is currently in fashion, with the concept of 'Networks of Excellence' pioneered in Canada in the late eighties (Fisher et al., 2001) and being implemented, for example, in the 6th Framework program of the European Union. The skewness of citation distributions is expected to give robustness to corresponding measures of excellence, this notion however being open to challenge when we consider changes in the reference set used for observation. In essence, an outstanding paper in a micro-community may get only a modest score when assessed within a larger field if the rest of this field has more generous referencing practices. At what extent cross-field differences, combined to an embedded classification scheme, create cross-scale differences that matter for performance evaluation?

Section I below summarises the data and methods, Sections II addresses the global stability of citation rankings at the various levels of aggregation / normalisation and Section III is devoted to the stability of 'excellence' based on top-cited articles. The Discussion and Conclusion sections then follow.

I -Data and Methods

1.1. Macro-structure

The present 'macro' approach is based on the following levels of observation: journal; specialities (ISI subject categories); sub-disciplines; disciplines -- the latter two are specific aggregates, by OST, of ISI subject categories. However, we have used a modified form of this structure, in order to obtain strictly embedded levels i.e. without overlaps¹. This strictly hierarchical structure is not of course the best model of the organisation of science, where disciplinary overlaps are a common feature, but rather a convenient simplification for the present study. Overlapping classes result in several normalised impact measures for each multi-assigned journal, which would obscure the proposed stylised analysis.

The five levels for this macro-study are defined as follows: L1: all science, L2: large-discipline (9 groups including multidisciplinary), L3: sub-disciplines (31 groups), L4: speciality (155 groups), L5: journal (3702 groups reduced to 3529 after filtering based on document type). The table is given in the Annex.es

In the following, the words 'field' and 'domain' are, here, general terms used as equivalents, whatever the scale. Groups at a particular level are termed as above:

- e.g. the term 'speciality' (level 4) corresponds, more or less, to the ISI "subject category" used in several Thomson-ISI products (SCI, JCR, WoS), with the afore-mentioned difference in relation to journals assignment: here we have assigned all journals to a single speciality.

- terms such as "discipline" may be misleading. For example, "civil engineering" may be held as a "discipline" by scientists in this area. As it is a "subject category" in ISI classification, we consider it as a speciality, level L4. The fact that ISI sometimes calls "discipline" the subject category may be confusing. For this reason, we do not use herein the term discipline alone, but only "large-discipline" to designate broad academic fields such as physics, fundamental biology, earth & space sciences, etc. The structure of the classification we have used is shown in Annex 1.

- it may be the case that a particular field survives through several levels of aggregation. This is due to the choices adopted for OST classification. One example is "mathematics", which, with the same journal set, is both a sub-discipline and a large-discipline (L3 and L2). But it splits at the L4 level. Another example is the field "multidisciplinary" which is a subject category defined by ISI, a quite heterogeneous grouping. Our classification keeps it as such, at the sub-discipline and the large-discipline level, there being no particular reason to join this group with any other.

Other variants of ISI-based classifications are found in literature, and the particular scheme we are using can certainly be questioned on a number of grounds. The classification we use, and this is true for the comparable variants, does not claim to be an accurate representation of the structure of science. It is, rather, a limited but "realistic" tool suitable for macro-analysis purposes. It is unlikely that using any other sensible "macro-classification" would profoundly alter the conclusions.

¹ To begin with we forced journals into a single speciality (subject category) using a random algorithm trying to avoid extreme losses in categories, but the 'specialities' in the present exercise do have fewer journals than the corresponding ISI subject categories. Then we modified the contour of OST sub-disciplines to a collection of specialities. In this strictly embedded scheme, a journal belongs to a single speciality, a speciality to a single sub-discipline, etc. The grouping of category codes (specialities) into the large academic disciplines we refer to is found in the annex of the biannual OST report, see for example Barré, Esterle (2002).

1.2 the empirical set (observed SCI) and two fictitious benchmarks

empirical SCI data

This research is based on primary data from ISI. Since the main aim is methodological, standard SCI coverage has been chosen. The publication year 1998 was selected, to give a sufficient delay for citations (i.e. a 4-year window). As far as the type of document is concerned, we wished to address a homogeneous population, avoiding for example the inclusion of 'review articles'. We therefore considered only the type 'article' (whether or not it came from proceedings or other sources). The empirical set makes use of real data and of a realistic structure of science.

In addition to this observed set, we used two fictitious sets which represent extreme models of citation distribution in order to benchmark the 'observed' set against a 'worst case' and 'best case' scenario, from the point of view of the cross-scale stability of indicators. To build these two sets, we have kept the same embedded structures L1-L5 without changing the size of classes (journals, specialities, etc.) at each level. But instead of their own real citations, articles were assigned new citation figures generated by a redistribution of original figures over the whole set, using two contrasting rules:

1st benchmark: 'ordered fields model'

This set was obtained using a redistribution of real scores between articles, by juxtaposing the real field structure and the list of ranked citation scores. As a result the first large-discipline in the list (conventional) gathers all the top-ranked articles, the second one the next highly cited articles, etc.; similarly, the 1st sub-discipline in the 1st large-discipline gathers the most cited articles in the large-discipline, and so on. The level of citation is forced to be completely dependent on the domain, at all levels of aggregation.

In this model, the respective proximities of two articles in terms of topic and of citation score are strongly dependent. The structure of science, reflected in the classification, dictates the citation behaviour. The expected result of this model is that normalised indicators, for a given article, will exhibit a very poor stability when the scale changes. For example, if the average citation of the domain is used as a normalisation factor, this factor will be different at the five levels considered (journal; speciality; sub-discipline; large-discipline; all). The model has no sampling rationale whatever the level. It represents the 'worst case' for the stability of citation indicators as the scale of observation changes. An interesting aspect of the ordered-field model is the supply of lower bounds for inter-level correlation values (see below).

2nd benchmark: 'random model'

For the second model the embedded structure (L1-L5) is also retained but with a random redistribution of the real citations scores amongst the articles: each article randomly receives the score of another one (we limited ourselves to one draw).

In this model the respective proximities of two articles in terms of topic (as described by the classification scheme) and of citation score are independent. The structure of science, operationalised by the classification, does not influence the citation scores. No field-dependence of citations is expected and each group in a classification, whatever the level (journal; speciality, etc.), will behave as a random sample of the overall distribution, reflecting the dispersion of citation scores of all science. Hence a sensible normalisation factor, say the average citation score of the domain, would tend to be stable through all the levels, as well as throughout the fields at any given level. We will illustrate below that the random model

implemented confirms this sampling rationale, though, theoretically, alteration of sampling features could occur in such strongly concentrated distributions (HAITUN, 1982). We refer to this model as the 'random model', expected to provide the 'best case' for the stability of normalised citation indicators as the scale changes.

The contents of the three tables, SCI, ordered fields and random model are too large to be displayed. A worked example, based on a (fictitious) miniature set of five small journals and two specialities, may be found in Annex 2.

The particular position of the 'observed' set, between the extremes, i.e. the 'best case' and 'worst case' configurations, gives an idea of the citation structure of 'real science' associated with the particular classification in use. Other benchmark models might be built with different properties. For example, a model with a strict hierarchy at the L5 level (with journals strictly associated with a prestige level) but a random mix of journals at the L4 level will tend to yield a zero correlation between L4 and L5, but a correlation close to one between all pairs of levels from L4 to L1. More generally, if at a given level fields are a random mix of sub-fields at the next lower level, then all fields at higher aggregation levels will maintain this sampling rationale. At the opposite end of the spectrum, the 'ordered fields' model above has no sampling rationale whatever the level. More sophisticated models could be used to try to emulate the properties of the observed SCI dataset. In Mandelbrot's wake, power-law or quasi power-laws encountered in bibliometrics have been interpreted in terms of self-similarity and put down to self-organisation mechanisms in science (for example Katz, 1999, Van Raan, 2000).

1.3. normalised citations: a rank approach

There are several ways to define field-normalised impacts: regular standardisation of variables; ratio to the field-average, which is commonly used; ratio to the field-average with log-transformation; and non-parametric positioning on ranks. Cardinal measures are more appropriate for some types of analysis, for example based on variance, after proper transformation to deal with the skewness of distributions. Distribution-free approaches may also be useful in addressing the problem in a general way. In this work we have privileged this 'ranking' approach, which also readily applies to 'excellence' measures. The rationale for rank comparisons is closer to the standardised-variables approach (the dispersion within each group is neutralised in both cases) than to normalisation using ratios to a central value (which keeps information on the dispersion of individual groups).

The principle is to rank all world articles, using quantiles on the criterion of citation score. A quasi-normalisation is obtained by using a local ranking at the level chosen: by discipline, by sub-discipline, etc. This method is very useful for positioning an actor by its 'activity index' profile in successive quantiles of citations. Fig 1 gives an illustration for a particular actor XXX (real case), with two profiles corresponding to discipline-level and speciality-level 'normalisations'. Abscissas correspond to classes of visibility, based on variable quantiles: 5% top-cited (excellence class), next 5%, next 10%, next 20%, next 20%, last 40%. Ordinates represent the activity index: for example the index would be 1.5 in the excellence class if 7.5% of the actors' publications fall into this class. Descending profiles correspond to a sound distribution of citations, the actor being more present in highly cited fractions. The figure shows that for this particular actor, changing the level of observation/normalisation alters the profile, see for example the activity index in the top-5% 'excellence class'.

In this study we do not focus on actors' comparison but on global differences on rankings over all SCI. We address the cross-scale stability using three entries, the overall correlation of rankings of all articles between the aggregation levels; the individual trajectory of articles throughout the levels; and the survival of the 'excellence' top-cited sets throughout the levels.

Overall comparison of article rankings

At each level of observation, L1 (all science) through L5 (journal)², articles are ordered by the citations they receive in the field they belong to, which puts them in particular quantiles (see footnote for technical issues³). L1 corresponds to the absence of field-normalisation: articles just receive their original ranking in all science (with for example a low expected performance of mathematics as a whole). In contrast, at the L4 level for example, the first percentiles of the overall ranking will be made of the most highly-ranked articles in each speciality. This corresponds to a strong field-normalisation. A step further, the first percentiles of the L5 overall ranking will contain the most highly ranked articles in each journal. Clearly, the percentile position of an article may change with the scale. For example a mathematical article may receive a mediocre score at the journal level L5, a good score at the speciality level L4 (if the journal has a good impact), and again a mediocre score at the L1 level, where mathematics are not favoured.

However, there are several built-in limitations to the discrepancies between levels. In the embedded structure L1-L5, the order relationship between two articles A and B belonging to the same field, at a given level, holds at all higher levels of aggregation. For example if article A is more cited than article B belonging to the same journal, this inequality is also true at the speciality level, the sub-discipline level, and all superior levels. This constraint mechanically bounds the discrepancies between rankings at different levels. Another mechanic factor that limits these discrepancies is due to the particular classification scheme adopted. As mentioned earlier the 'multidisciplinary' field persists over three levels (large-discipline, sub-discipline, speciality) and 'mathematics' over two levels (large-discipline, sub-discipline). An advantage of the 'ordered-fields' model is to give a measure of the lower bounds of correlation between levels due to these built-in factors.

Another factor increases somewhat artificially the correlation between rankings at different levels, that is the nature of citation distributions, which creates a huge number of ties in low score classes (particularly zero or one citation). In order to reduce this effect, we used truncated sets (without the zero citation articles; without the less cited half of all articles⁴) together with the original set with all articles.

The global agreement between rankings of all articles between all (pairs of) levels was assessed by Kendall's rank correlation, based on the difference of concordances and discordances of ranks for each pair of items, of direct interest here. We tested several quantiles grids (20 half-deciles, 100 percentiles, 500 and 1000 quantiles), and the results tend to converge as the grids become finer and finer, the difference between grids '500' and '1000' no longer affects the last digit in the subsequent tables⁵. Results (in the form of 5 by 5 tables) are therefore reported using grid '1000'. For the reason stated above

² L6 (document level) is degenerated, clusters are reduced to individual documents.

³ The different frequencies by field / level obliges us to use quantiles rather than citation scores. Let us take the example of a 100 cells grid (percentiles). Thus article i belongs to percentile x_i in its journal, to percentile y_i in its speciality, and so on. The skewed nature of the distribution of citations creates multiple ties for lower scores, e.g. zero or one citation. Ranking procedures offer several options for ties i.e. low, average, high, or random ties. For the global analysis we used the 'high ties' option which gives the least favourable rank. Thus all zero citations articles belong to the same last class with the lowest rank (e.g. 100 for centiles). As a result the size of low percentile classes may be irregular, which is true for all tie options except 'random ties'. The "random ties" option is recommended for comparisons with the two extreme models.

⁴ Truncation has been carried out at the speciality level. Because of ties, only 45% of the articles were kept, representing almost 9/10 of the total citations.

⁵ The choice of a unique grid applicable to all levels is a trade-off. If the grid is 'coarse-grain' (say percentiles), a large number of ties is created at the higher levels of aggregation, and the rank comparisons between levels tend to give higher rank correlation than the fine-grain grids. If a fine-grain (say 10,000) is chosen, many gaps are created at the lower level(s) of aggregation. However statistical packages can handle this problem.

(persistence of order relations in couples when the aggregation level increases) it may be useful to judge the difference to independence by considering the lower boundary values given by the tables 'ordered fields model' as a reference, instead of the zero correlation.

Individual trajectory of articles

Thereafter, a study of individual article's trajectories, between L1 through L5, is sketched out. One simple measure of the total trajectory length is $\sum_{j=1..4} [| x_{i(j+1)} - x_{ij} |]$. This index measures the degree of change in scores as the level of aggregation changes within the embedded structure adopted. Another interesting feature is the range of variation $\text{Max}_{j=1..4}(x_{ij}) - \text{Min}_{j=1..4}(x_{ij})$. Distributions of the trajectories' lengths and ranges give an idea of the stability of scores throughout the levels.

Excellence measures using top-cited articles

Among other indicators the contents of the 'top end' of bibliometric distributions (and especially citations) is a typical way in which 'excellence' is measured. An important question is whether such a measure can resist changes of scale. We chose three operational definitions for the excellence class, ranging from a 'looser' to a 'tighter' definition:

- the 1st half-decile of the impact distribution, i.e. using grid: '20'; ca. 25,000 top-cited articles for all SCI -
- the 1st percentile, grid: '100'; ca. 5000 top-cited articles
- the first 0.2% (grid: '500'; ca. 1000 top-cited articles)

For each level of aggregation L1-L5, and with each given threshold, the top-cited set was built. By construction, at any level, the number of 'excellent' items from a particular field is proportional to the total size of this field – with respect to the fluctuations due to ties. For ranking, the 'low ties' option, giving the most favourable rank, was selected to pick up at least one article in each journal. Obviously not all grids can be applied to the whole range of levels, in order to respect a minimum size of selected sets. The grids '100' and '500' (in particular) are meaningless at the journal level L5, which contains ca. 3500 journals; the grid '20' is already questionable at this level; so is the grid '500' at the speciality level.

The question then is "do the contents of the top-cited class - in terms of articles - change when the level of aggregation / observation changes"? We addressed this question by studying:

- the degree of overlap of 'top-cited sets' between each pair of levels
- the degree of global overlap between top-cited sets at all levels, giving the persistence of a top-cited class.
- the set of articles considered top-cited at (at least) one level.

II - Are citation scores stable across-scale?

Impacts, a core topic in bibliometrics, have received a great deal of attention in the literature. The 'state-of-the-art' concerning impact factors was discussed in particular by Glaenzel & Moed (2002). Recent debates about impact factor measures have been commented upon by Moed (2002) and the implementation of relative citation rate analyses using a normalisation at the journal level were carried out at the ISSRU group (Schubert & Braun, 1986). The literature is full of applications of normalisation at various levels of aggregation, including micro-levels. For example, normalisation at the level of small clusters of journals was used by Bassecoulard & Zitt (1999) to select local cores of journals; normalisation at the level of small groups of papers by Kostoff in his studies of team or researcher performance (2002, see also Vinkler, 2002). The perennial question of field-normalisation was further reviewed by Schubert & Braun (1996 op.cit.). Also, general views related to relative indicators have been developed by for example Glaenzel, 1988 and Egghe & Rousseau, 2003. Most producers of indicators have for decades offered relative measures (i.e. relative impact measures: citation share over publication share) or ranking

measures which make comparisons between fields possible at a given disaggregation level. The properties of various normalisation techniques have to be explained when a measure at a given level is intended to reflect structural discrepancies at lower levels. In reaction to the common practice among decision-makers, using straightforward 'ISI impact factors', the flow of articles recommending various forms of field-level disaggregation and/or normalisation is constant (see for example amongst others Sen, 1992, Marshakova-Shaikevich, 1996, Csapski, 1997, Ramirez et al., 2000, Solari & Magri, 2000). As mentioned above, we wish to address here a cross-scale perspective adding further to this cross-sectional perspective.

2.1. The general landscape

The differences in average impact between ISI categories are well-known. For the data and the particular delineation of specialities used here, using four years citation windows, the general picture on observed SCI is as follows:

Table 1 - average impacts (4-years window) - examples of field discrepancies across three levels

	L2 (9 large disciplines)	L3 (31 sub-disciplines)	L4 (155 specialities)
fields with high values <u>max value (underlined)</u>	<u>multidisc</u> fundamental biology, medical research	<u>multidisciplinary</u> bio/cellchem immunology oncology, astronomy	<u>multidisciplinary</u> embryology mol.biology immunology biochemistry
examples of fields with median values	earth &space physics chemistry	general physics health chemistry (narrow) pharmacy applied physics	parasitology limnology dermatology microscopy surgery
fields with low values <u>min value (underlined)</u>	applied biology engineering, <u>mathematics</u>	food science materials computer/info mech/fluid eng. <u>mathematics</u>	geotech. mine engin. mater./analysis civil engin. <u>photography</u>
Across-field ratio: max/min	11	11	30

The across-field ratio is mechanically equal for L2 and L3 because the conventional classification adopted maintains the categories 'multidisciplinary' (highest impact) and 'mathematics' (lowest impact), unchanged, at L2 and L3 levels. The increase of the ratio max/min for L4 is due to the poor citation performance of a small speciality.

Figure 2 shows the distribution (arithmetic mean, weighted) of fields' mean impact at the L2 through L5 levels. Analysis of the geometric mean and first quartile distribution (not shown) lead to similar results. The reduction of variance with aggregation is lower than expected, showing that discrepancies in citation behaviour are maintained through a large range of scale. The picture would be completely different in the 'random model'.

2.2. stability of rankings in scale changes

As explained in the methodological section, we evaluated the level of discrepancy between citation rankings of articles for each combination of levels, for the three cases: the observed set (empirical SCI data); the random model; and the 'ordered fields' model. If the structure of science embodied in the classification large-discipline/ sub-discipline/ speciality/ journal were based on random mixes (random model), a quasi-maximum correlation would be expected for all pairs of levels. As mentioned above, this

would be the 'best case' for the cross-scale stability of citation indicator, the score of articles would be practically unchanged whatever the level of observation. In contrast, the worst case corresponds to a structure of science based on a strictly 'ordered fields' model, with very low correlation indices expected. Tables 2A-E show the Kendall rank correlation between ratings at different levels.

Table 2-A - global rank correlation between levels - observed SCI - grid 1000 - ties option: high

Kendall tau grid 1000	L1 (all science)	L2 (large disc.)	L3 (sub-disc.)	L4 (speciality)	L5 (journal)
L1	1.00	0.81	0.76	0.72	0.51
L2	0.87	1.00	0.86	0.80	0.54
L3	0.84	0.91	1.00	0.86	0.56
L4	0.82	0.87	0.91	1.00	0.58
L5	0.66	0.68	0.69	0.71	1.00

for all Tables 2 except 2E:

above diagonal: articles with zero citation excluded

below diagonal: all articles

Table 2-B - global rank correlation between levels - observed SCI - grid 1000 - ties option: random

Kendall tau grid 1000	L1 (all science)	L2 (large disc.)	L3 (sub-disc.)	L4 (speciality)	L5 (journal)
L1	1.00	0.77	0.71	0.67	0.44
L2	0.83	1.00	0.85	0.78	0.49
L3	0.79	0.89	1.00	0.85	0.52
L4	0.76	0.83	0.88	1.00	0.55
L5	0.55	0.59	0.61	0.63	1.00

Table 2-C - global rank correlation - random model - grid 1000 - ties option: random

Kendall tau grid 1000	L1 (all science)	L2 (large disc.)	L3 (sub-disc.)	L4 (speciality)	L5 (journal)
L1	1.00	1.00	0.99	0.99	0.94
L2	1.00	1.00	1.00	0.99	0.94
L3	0.99	1.00	1.00	0.99	0.94
L4	0.99	0.99	0.99	1.00	0.94
L5	0.94	0.94	0.94	0.94	1.00

Table 2-D - global rank correlation between levels - 'ordered fields' model - grid 1000 - ties option: random

Kendall tau grid 1000	L1 (all science)	L2 (large disc.)	L3 (sub-disc.)	L4 (speciality)	L5 (journal)
L1	1.00	0.18	0.04	0.01	0.00
L2	0.17	1.00	0.22	0.08	0.00
L3	0.04	0.23	1.00	0.27	0.02
L4	0.01	0.09	0.27	1.00	0.06
L5	0.00	0.01	0.02	0.06	1.00

Table 2-E - global rank correlation between levels - observed SCI (truncated) - grid 1000 - ties option: random

Kendall tau grid 1000	L1 (all science)	L2 (large disc.)	L3 (sub-disc.)	L4 (speciality)	L5 (journal)
L1	1.00	0.71	0.64	0.57	0.50
L2		1.00	0.80	0.69	0.59
L3			1.00	0.79	0.65
L4				1.00	0.74
L5					1.00

data for truncated SCI: only the 50% most cited articles (selection at the speciality level) are kept

Tables 2A, 2B, 2E show results in relation to the 'real' (observed) SCI. whereas tables 2C and 2D refer to the two 'benchmark' sets.

Rank correlation tends to decrease, as expected, with the distance between levels. For example, in table 2A depicting the observed SCI, on the row L1, the coefficients decrease from column L2 through column L5. The correlations between adjacent levels remain fairly high (>0.80, table 2A above diagonal), but are far from expressing a sampling situation as described in the random model table 2C. The striking contrast with the 'ordered fields' model (2D) with its absolute hierarchy and its very low correlations is less surprising. Over four levels (all science vs. speciality), the correlation drops to 0.72.

The journal level is a particular case. Compared with all others, this level shows a strong discrepancy, with correlation coefficients of only 0.6 with the next level L4 (Table 2A, figures without zero citation items). This is only partly due to the inter-level distance, since on average the speciality has a size >20 times a journal, roughly the same factor as between L2 and L4, but correlation L2-L4 is much higher than L4-L5. The singular nature of the journal level supports the idea of a twofold competition: for accessing the best journals on the one hand; and for gaining visibility within each journal on the other hand. The corresponding classical indicators are respectively the Expected Impact (or actor's impact factor) and the Relative Citation Ratio (RCR, Schubert & Braun, op.cit. 1986). The global impact can be decomposed as the product of Expected Impact and RCR. A two-step model of between-journal and within-journal competition has been proposed by Van Raan (2001).

The singular nature of the journal level suggests that the real model of science is a mixed one. The passage L5-L4 is consistent with a mixing process of journals strongly unequal (L5), but only to a certain extent. As a result, the aggregation brought by the speciality level reduces the variety, but not all the variety. If specialities were really based on a random sampling of journals, we would have a mixed scheme, with a very low L5-L4 correlation coefficient, and all the subsequent coefficients (all pairs of levels among L1-L4) close to one. In such a situation, all citation scores could be considered as stable for all aggregation levels beyond L5. This is not the case, and levels from L4 through L2, in the classification adopted, always retain some inter-field variety, and as a result, a cross-scale instability of citation rankings.

The correlation indices are sensitive to the removal of uncited or little cited articles. Correlation for complete sets may be considered spurious because of the abundance of ties for low scores. Light truncation (discarding zero-citation articles) and strong truncation (only keeping the half more cited, table

2E) make clearer the instability of citation rankings across scales. In table 2e, we watch the effect for the large discipline level normalisation, with a correlation coefficient of rankings L1 and L2 scarcely above 0.7. We will consider further, in section III, the effects of a drastic truncation which retains only the top of the distribution.

2.3. individual trajectory length

Low Kendall coefficients indicate a global disarray between scores of two particular levels. A complementary view is provided by looking at trajectories of individual articles from L1 through L5 levels, using indices defined in § I.3. If the trajectories are short, it is an indication that the structure of science adopted does not affect individual article benchmarking. In this case the organisation would be close to the random model. Long trajectories suggest that the choice of the aggregation level in the structure of science heavily affects the citation score and subsequent evaluations. This would be the case for the 'ordered fields' model for example.

The observed SCI data, on the 1000-grid, yield an average total trajectory of ca. 300 positions in the four intervals L1 through L5, i.e. about 75 positions (0.75 decile) at one level change. The distribution of the trajectory length for the observed SCI set is shown in Fig.3. The standard deviation of the total trajectory is about 165 positions, and the maximum reaches 1125, i.e. 280 positions (almost 3 deciles) for one level change. The long tail indicates the presence of strong deviations (log-normal shape).

III - Top-cited articles

There are many ways to approach 'excellence' by bibliometric measures, for example by considering the tails of bibliometric distributions (production, collaboration, citation, etc.), the presence in particular categories (funding and scientific committees), the position in various networks (strategic themes, collaborations, etc.). Highly-cited articles are among the most commonly used indicators (see for example Garfield, 1986 on the very top articles). Indicator producers generally trust this type of measure but at the same time highlight the technical caveats and the necessity for corroborative results from several points of view (Tijssen et al., 2002). Studies by sociologists and bibliometricians on citation analyses have brought evidence that citations are a marker of communication and visibility, but the danger of over-interpreting them as a marker of quality is a threat to accurate research evaluation (see amongst many others Seglen, 1997, McRoberts & McRoberts 1989; and for the opposing argument, see for example Vinkler, 2004). This is particularly true for measures of excellence. We do not wish to address here the issue of fundamental limits of citations for measuring excellence, but the simple technical stability of top-citedness indicators as the scale of normalisation changes.

The rationale is the same as in the previous sections, with a simple change in focus to the top-end of the distribution: we considered three selections, the very top class (0.2%) designated as E1, the first percentile (1%) as E2, and the first half-decile (5%) as E3.

Tables 3 A-C show the overlaps between pairs of levels. For example, the first figure in the box L1-L2 (Table 3A) indicates that for the very top class, the overlap between this class defined at the L1 level (all science) and at the L2 level is 52%.

Table 3-A - first class= 0.2% of publications - overlap (%) - observed SCI - ties option: low

Overlap (%) for E1 (first 0.2%)	L1 (all science)	L2 (large disc.)	L3 (sub-disc.)	L4 (speciality)	L5 (journal)
L1		52.0	38.6	<i>33.1</i>	ns
L2	51.6		64.5	<i>50.9</i>	ns
L3	38.0	63.9		<i>69.5</i>	ns
L4	<i>30.7</i>	<i>47.5</i>	<i>65.5</i>		ns
L5	ns	ns	ns	ns	

all tables 3A-C:

above diagonal: as a fraction of the top-cited class at higher level

below diagonal: as a fraction of the top-cited class at lower level (the two fractions usually differ because of ties)

italics: small average size of units

Table 3-B - first class= 1% of publications - overlap (%) - observed SCI - ties option: low

Overlap (%) for E2 (first percentile)	L1 (all science)	L2 (large disc.)	L3 (sub-disc.)	L4 (speciality)	L5 (journal)
L1		60.7	48.7	42.2	ns
L2	60.2		71.8	60.8	ns
L3	47.6	70.8		71.7	ns
L4	40.7	59.2	70.7		ns
L5	ns	Ns	ns	ns	

Table 3-C - first class= 5% of publications - overlap (%) - observed SCI - ties option: low

Overlap (%) for E3 (first half-decile)	L1 (all science)	L2 (large disc.)	L3 (sub-disc.)	L4 (speciality)	L5 (journal)
L1		70.4	64.3	58.1	<i>39.4</i>
L2	70.1		80.6	72.7	<i>45.6</i>
L3	63.2	79.7		80.2	<i>49.1</i>
L4	56.9	71.5	79.8		<i>54.1</i>
L5	<i>31.8</i>	<i>36.9</i>	<i>40.2</i>	<i>44.6</i>	

The tables above show that the overlaps between excellence classes defined at different levels are only moderate, and tend to decrease as the threshold of excellence goes up. For example, for the 1% excellence level, the overlap between the all-science excellence class and sub-discipline or speciality-excellence class is less than 50%. Between two candidate levels for practical normalisation, L3 and L4 (L5 having a clearly different meaning), the overlap is only of two thirds. The benchmark models, not shown, behave as expected: overlaps close to 100% for the random model, and very weak for the 'ordered fields' model.

Let us now consider the across-level stability in a wide range of levels: L1 through L4, leaving aside the journal level, and L1 through L5 including the journal level only when it is not meaningless. The question is "which fraction of articles is top-cited whatever the level?" i.e. constantly retrieved in the top-cited class (Table 4A).

Table 4-A - constantly top-cited articles - observed SCI - ties option: low

level of 'excellence'			Frequency		% of total publications bold: (% of nominal size of top-cited class)	
grid level	percentage of total publications	Nominal size of top-cited class (thousands)	L1thr.L5	L1thr.L4	L1thr.L5	L1thr.L4
E1: 500	0.2%	1.1	ns	0.3	ns	0.06% (27%)
E2: 100	1%	5.6	ns	2.0	ns	0.36% (36%)
E3: 20	5%	27.8	8.6	14.7	1.55% (31%)	2.64% (53%)
total	556.8	556.8	556.8	556.8	100.00	100.00

Table 4A shows that the fraction of top-cited papers remaining top-cited is low, except for E3 in the window L1-L4 where it reaches 50%. For example, if we use the first percentile as defining top-citations, the total overlap between the top-classes defined at four levels L1-L4 covers only one third (36%) of a percentile. Generally speaking, only a minority of top-cited articles resist the scale change. The overlap becomes much lower, as expected, when the journal level is included.

From another perspective, the information about "articles being top-cited at least at one of the five levels" is given in table 4B.

Table 4-B - articles top-cited at least at one level - observed SCI- ties option: low

level of 'excellence'			Frequency		% of total publications bold: (% of nominal size of top-cited class)	
Grid level	Percentage of publications pub	nominal size of exc. class (thousands)	L1thr.L5	L1thr.L4	L1thr.L5	L1thr.L4
E1: 500	0.2%	1.1	Ns	2.4	ns	0.43% (215%)
E2: 100	1%	5.6	Ns	10.6	ns	1.91% (191%)
E3: 20	5%	27.8	62.9	45.9	11.3% (226%)	8.25% (165%)
total	556.8	556.8	556.8	556.8	100.00	100.00

The excellence class is roughly doubled with respect to the nominal threshold, except again for E3/L1-L4. For example if we define the top-cited class as the first percentile (row 'E2') we find about two percentiles of all scientific publications that can be considered as excellent somewhere. The benchmark models, not shown, exhibit the expected results: the persistence of the top-cited classes is very low for the 'ordered fields' model, and close to 100% for the random model.

These empirical findings confirm that the contents of 'top-cited classes' are strikingly dependent on the level of aggregation/ normalisation. Secondly, the contents of the top-class are all the more unstable as the selection of the 'elite' is stringent. This effect may be linked to fact that in large top-cited sets, a long trajectory (fig.3) is needed for the fraction of most-cited items to escape the set when the level of aggregation changes. The low probability of these long trajectories, compared to medium ones, could enhance the stability of large sets compared to very highly cited ones.

main results

It is widely recognised that the diversity of citation practices in scientific fields justifies some form of field-normalisation. The field-normalisation here was operationalized by a local quantile score for each article in its field. But scientific publications, for example the SCI set, can be aggregated at various levels, e.g. from a 'macro' perspective: journals, specialities, sub-disciplines, disciplines, and all science. Depending on the level, the quantile score of an article is subject to change. It would be substantially the same for cardinal measures, for example citations normalised by a central field value. As a result, field-normalised indicators are not only, trivially, dependent on the delineation of fields, but also, for a given multi-level classification, dependent on the hierarchical level of observation in a particular classification. An article may exhibit very different citation scores or rankings when compared within a narrow speciality or a large academic discipline. More generally, evaluation based on citation scores is likely to be wholly dependent on the level of aggregation used. This is not a purely theoretical debate. In universities or research institutes, especially if they cover a wide scope of fields, the question unavoidably arises of the breakdown level.

We have studied, empirically, the stability of citation scores using one year of SCI data and a conventional classification scheme, realistic from the point of view of macro-level indicators of science. We have considered five levels of observation / normalisation: all science (e.g. original rankings); large academic discipline; sub-discipline; speciality; and journal. Certain limitations of the methodology should be borne in mind:

- the science classification we have used may be called into question, including the composition of ISI subject categories. The strict embedding of the structures we have adopted for the sake of simplicity (one article has a single trajectory) may also exaggerate some irregularities that would be smoothed by multiple assignment, especially at the journal level.
- a cardinal approach instead of rankings, using variance analysis at each level with an appropriate transformation of variables, might also be a helpful complement.
- as far as excellence is concerned, the analysis was carried out on articles. For the benchmarking of authors or institutions, aggregates may be more robust.

However, these issues would be unlikely to alter the main findings:

1- We investigated the citation scores, in terms of relative ranking, at various levels of observation, in a realistic field-classification scheme. The variation of citation scores depends on the particular distribution of articles citations vis-à-vis the classification scheme. We observed a large degree of discrepancy of citation rankings among levels of observation.

2. The journal level is a very particular case, and differs greatly, in all cases, from all other levels. This phenomenon is similar to discrepancies observed by scholars, using cardinal measures, between 'relative citation ratio (RCR)' at the journal level and overall impact measures. Furthermore, compared to the more extreme benchmark models, 'real science' depicted here by an actual citation distribution among articles, appears as a mixed model, with a large but partial mixing process taking place when aggregating journals into specialities. Further aggregation keeps conveying discrepancies in terms of citation behaviour and, as a result, instability of citation indicators.

3. The distribution of individual trajectories of articles' rankings over all the levels has a long tail, indicating that for particular articles the discrepancies between ratings at different levels are severe. A similar observation could be anticipated from the average values of impact at each node of the tree corresponding to the particular classification used.

4. Turning to the top-cited fractions, a classic measure of excellence, it becomes clear that a large proportion of the top-cited set does not survive the change of scale. Moreover, for sensible levels of definition of the top class, the narrower the definition of excellence, the greater the instability.

The fact that citation indicators are not stable from a cross-scale perspective is a serious worry for bibliometric benchmarking. What can appear technically as a 'lack of robustness' raises deeper questions about the legitimacy of particular scales of observation. A few questions are particularly appealing for future research:

the need for a micro-level approach

We used a macro-classification, a realistic but perhaps limited approach. A purely theme-based "micro-approach", i.e. based on grouping of individual papers rather than journals, would perhaps be more suitable to address these theoretical issues of scientific structure. This method would bypass the 'journal' step which conveys a mixed logic of thematic delineation and prestige level. In the second stage of this study, devoted to the micro-approach, we intend to define successive levels using bibliometric groupings (research fronts, bibliographic coupling clusters, etc.) rather than the conventional breakdowns based on ISI subject categories. Because of computational constraints however, the investigation will perforce be limited to a particular field.

cited-side or citing-side normalisation?

In this paper we have addressed normalisation in a typically classical way, namely from the viewpoint of the cited side. This is perfectly adequate as long as the cited field and the citing field coincide. But knowledge transactions, and from Mertonian perspective their citation counterpart, take place between fields as well as within fields, whatever the level of analysis, and multi-disciplinarity is a key dimension of the evolution of research systems. In such cases, one should be aware of the fact that discrepancies in behaviour, and in particular in referencing practices, originate from the citing side. Together with other aspects (such as the regimes of growth), the length of the reference list, rather than the size of the field (on the latter point see Garfield, 1998) is responsible for the citation biases we would wish to neutralize. The question arises of a normalisation at the source rather than at the reception of citations, either by some fractional count of emitted citations, used for example by Zitt et al. (1994) in a co-citation context (Small, 1985, credits Thomson and Dean as the originators of this fractional option) or by a field-normalisation on the citing side.

strong normalisation vs. no normalisation and the case of mutlidisciplinary articles

There are a number of arguments for the "smallest set" strategy. The rationale is: the smaller the set, the more relevant the comparison. At the macro-level, the choice of the level of ISI subject categories, though purely arbitrary (and possibly with questionable delineation), is often made because it is simply the smallest basis widely available. Micro-approaches can provide still smaller clusters of articles. Kostoff, (op.cit. 2002) investigated normalisation at this level. The use of ISI research fronts or other micro-analysis clusters could also be a possible approach. The radical statement about the differentiation process of research by Wasmer (2001), who notes with Fitoussi that, in the extreme, "*any researcher is in a monopoly situation and can by definition not be evaluated*", illustrates the potential limit case.

But there also arguments against over-normalisation. One is purely technical, sets which are too small jeopardize statistical stability. The other argument is more fundamental and relates to trans-disciplinarity. Over-normalisation can unduly level down all specialities or themes in science, and under-rates the role of articles in leading or central topics. Examples of potentially leading articles are documents with a multidisciplinary scope and/or published in multidisciplinary journals. At the macro-level, the 'multidisciplinary' category of SCI, where Nature, Science and PNAS are found, has a very high citation average. If a crude category-level normalisation is applied, using for example its average citation as the normalisation factor, the visibility of many strong articles will be under-estimated. The same will be true for a micro-approach grouping multidisciplinary articles with similar profiles in the same cluster. The normalization on this reference will tend to under-rate their significance, while a 'quick and dirty' normalisation using their main disciplinary assignment, or some all-science mix, would produce better results in this particular case. Again citing-side normalisation may be helpful.

not one best level

A variety of points of view need to be accepted. This was already the implicit message of the promoters of RCR, coupling two levels of analysis, the relative citation ratio (local view) and the expected impact combined into the regular impact (global view). Kostoff (op.cit.) in another context talks about the combination of "job right" (local view) and "right job" in a global performance. It remains the case that the choice of the best level for delineating the topic or the job is largely arbitrary since scientific networks exhibit to a large extent a self-similar structure.

Not much can be expected from the analysis of local stability across levels. The persistence of an article in the top-cited class is an example. The rationale of such multi-level ratings is weak, many of the scale-resistant articles simply have the good fortune to belong to specialities which are, perhaps only as the result of citation habits, the most visible. In quasi-continuous classification schemes used in "micro-bibliometrics", local maxima of stability can also be looked for as the level of observation changes, but again the choice of a single level is questionable.

The absence of 'one best level' of observation has a particular consequence for excellence measures. In a relativist view, if all levels are equal in terms of legitimacy, each scientist can 'choose his/ her level of excellence' by picking the aggregate where his/her articles will score the highest in a sensible classification scheme.

Without going to such an extreme, exclusively thinking in terms of 'vertical bibliometrics' and not paying attention to variety and emergence may be a counter-productive approach. At the very least, vertical bibliometrics should cope with cross-scale instability and arbitrary choices of levels of observation, and offer shifting viewpoints with different zoom settings. This suggests a particular prudence in interpretation of citation indicators, including 'excellence' studies based on top-cited fractions.

Acknowledgements

The authors would like to thank Ronald Kostoff for comments on a first version of this text, and anonymous referees for helpful suggestions.

References

- BARRÉ R., ESTERLE L. (2002), *Science & technologie. Indicateurs 2002. Rapport de l'Observatoire des Sciences et des Techniques*, Economica, Paris.
- BASSECOULARD E., ZITT M. (1999), Indicators in a research institute: A multi-level classification of scientific journals, *Scientometrics*, 44 (3): 323-345.
- CZAPSKI G. (1997), The use of deciles of the citation impact to evaluate different fields of research in Israel, *Scientometrics*, 40 (3): 437-443.
- EGGHE L., ROUSSEAU R. (2002), A general frame-work for relative impact indicators, *Canadian Journal of Information and Library Science-Revue Canadienne Des Sciences De L Information Et De Bibliotheconomie*, 27 (1): 29-48.
- FISHER D., ATKINSON-GROSJEAN J., HOUSE D. (2001), Changes in academy/industry/state relations in Canada: The creation and development of the networks of centres of excellence, *Minerva*, 39 (3): 299-325.
- GARFIELD E. (1986), Do Nobel prizes winners write citation classics ?, *Current Contents*, (23): 182.
- GLAENZEL W., MOED H. F. (2002), Journal impact measures in bibliometric research, *Scientometrics*, 53 (2): 171-193.
- HAITUN S. D. (1982), Stationary scientometric distributions. Part ii. Non-gaussian nature of scientific activities, *Scientometrics*, 4 (2): 89-104.
- KATZ J.S. (1999). The self-similar science system, *Research Policy*, vol 28, n°5, pp. 501-517.
- KOSTOFF R. N. (2002), Citation analysis of research performer quality, *Scientometrics*, 53 (1): 49-71.
- MARSHAKOVA-SHAIKEVICH I. (1996), The standard impact factor as an evaluation tool of science fields and scientific journals, *Scientometrics*, 35 (2): 283-290.
- MC ROBERTS M. H., MC ROBERTS B. R. (1989), Problems of Citation Analysis: A Critical Review, *Journal of the American Society for Information Science*, 40 (5): 342-349.
- MOED H. F. (2002), The impact-factor debate: the ISI's uses and limits, *Nature*, 415 (14 Sep): 731-732.
- MURUGESAN P., MORAVCSIK M. J. (1978), Variation of the nature of citation measures with journal and scientific specialties, *Journal of the American Society for Information Science*, 29: 141-155.
- NARIN F. (1976), *Evaluative bibliometrics : the use of publication and citation analysis in the evaluation of scientific activity*, National Science Foundation, Contract NSF C-627,
- PINSKI G., NARIN F. (1976), Citation influence for journal aggregates of scientific publications : theory, with application to the literature of physics, *Information processing and management*, 12: 297-312.
- RAMIREZ A. M., GARCIA E. O., DELRIO J. A. (2000), Renormalized impact factor, *Scientometrics*, 47 (1): 3-9.
- SCHUBERT A., BRAUN T. (1986), Relative indicators and relational charts for comparative assessment of publication output and citation impact, *Scientometrics*, 9 (5-6): 281-291.
- SCHUBERT A., BRAUN T. (1996), Cross-field normalization of scientometric indicators, *Scientometrics*, 0 (0): 1-14.
- SCHUBERT A., GLAENZEL W., BRAUN T. (1988), Against absolute methods : relative scientometric indicators and relational charts as evaluation tools, In: A. F. J. VAN RAAN (Eds), *Handbook of Quantitative Studies of Science and Technology*, Elsevier, Amsterdam, pp. 137-169.
- SEGLEP P. O. (1997), Citations and journal impact factors: questionable indicators of research quality, *Allergy*, 52: 1050-1056.
- SEN B. K. (1992), Documentation Note Normalized Impact Factor, *Journal of Documentation*, 48 (3): 318-325.
- SIEGEL S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New-York (USA).
- SMALL H., SWEENEY E. (1985), Clustering the Science Citation Index using co-citations I - a comparison of methods, *Scientometrics*, 7 (3-6): 391-409.
- SOLARI A., MAGRI M. H. (2000), A new approach to the SCI Journal Citation Reports, a system for evaluating scientific journals, *Scientometrics*, 47 (3): 605-625.
- TIJSSSEN R. J. W., VISSER M. S., VAN LEEUWEN T. N. (2002), Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference?, *Scientometrics*, 54 (3): 381-397.
- VAN RAAN A. F. J. (2000). On growth, ageing, and fractal differentiation of science, *Scientometrics*, vol 47, n°2, pp. 347-362.
- VAN RAAN A. F. J. (2001), Competition amongst scientists for publication status: Toward a model of scientific publication and citation distributions, *Scientometrics*, 51 (1): 347-357.
- VINKLER P. (2002), Subfield problems in applying the Garfield (Impact) Factors in practice, *Scientometrics*, 53 (2): 267-279.
- VINKLER P. (2004), Characterization of the impact of sets of scientific papers: The Garfield (Impact) Factor, *Journal of the American Society for Information Science and Technology*, 55 (5): 431-435.
- WASMER E. (2001) Some political economy of excellence, *Workshop "In search of scientific excellence : research*

performance by discipline", EU, STI-ERA, Nov 13.

ZITT M., BASSECOULARD E. (1994), Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis, *Scientometrics*, 30 (1): 333-351.

ZITT M., RAMANANA-RAHARY S., BASSECOULARD E. (2003), Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation, *Scientometrics*, 56 (2): 259-282.

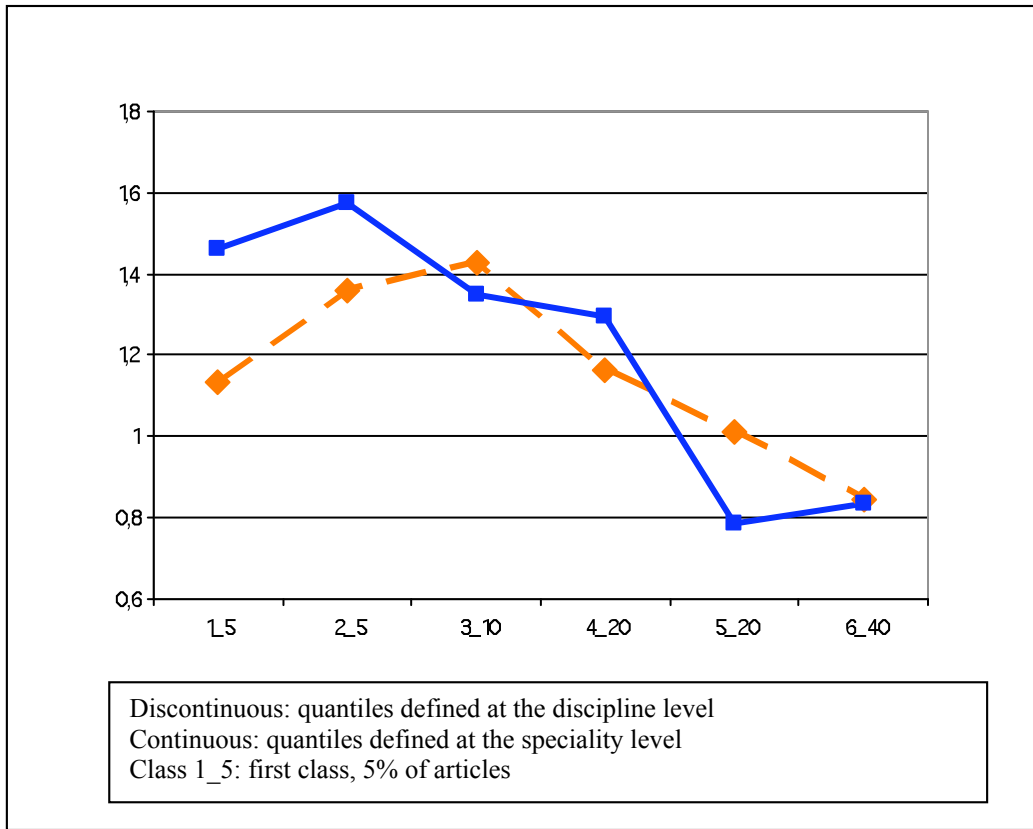


Fig. 1 ACTOR 'XXX': PROFILE USING ACTIVITY INDEX IN VISIBILITY CLASSES

In a given class, for example the first one defined as the top-cited 5% , noted 1_5, the activity index is the ratio of the percentage of the actors' articles falling in this class, to the percentage of the world articles in the class. Real values used in the denominator may be different from the nominal values (5%, 5%, 10%, etc.), especially for the last quantiles, because of ties.

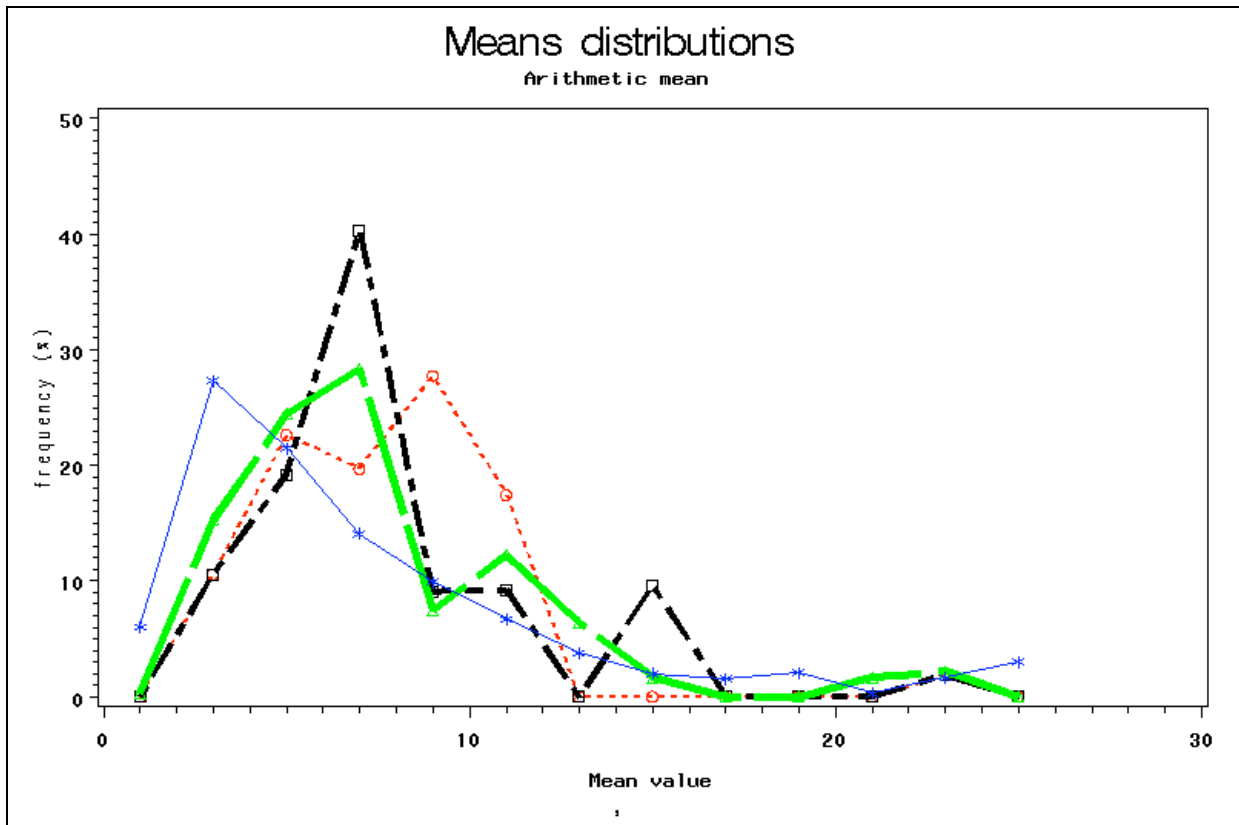


Fig. 2 DISTRIBUTION OF MEAN IMPACT BY FIELD (empirical SCI set)

triangles: fields=disciplines (L2)
squares: fields=sub-discipline (L3)
circles: fields=speciality (L4)
stars: fields=journals (L5)

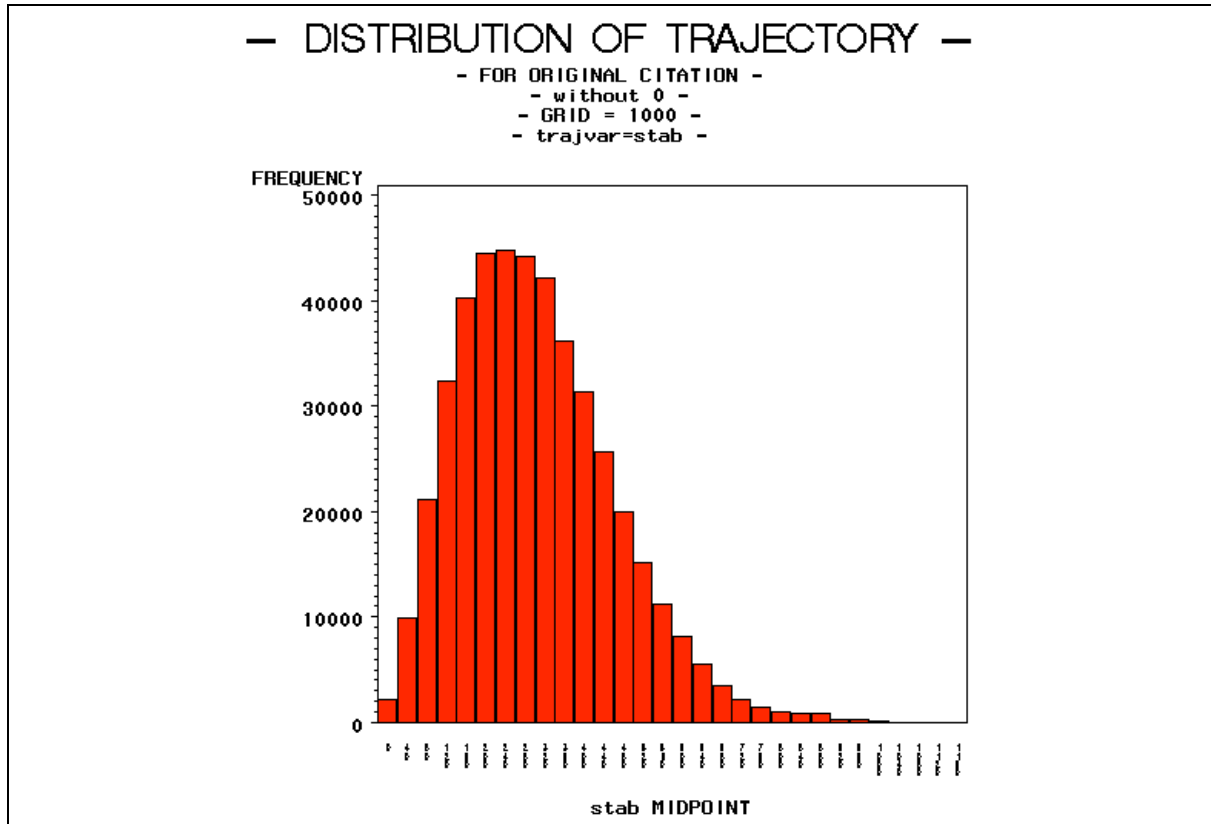


Fig. 3 TRAJECTORY LENGTHS DISTRIBUTION (empirical SCI set - grid=1000)

ANNEX 1

The table represents an extract of the classification used, for two large disciplines. As mentioned, 'specialities' are based on ISI 'subject categories' but for the need of this particular exercise journals are assigned to a single one, which can make a large difference in their contents. Sub-disciplines as sets of specialities are adapted from OST sub-disciplines, again without multi-assignment. Large academic disciplines are those used at OST.

disc	discipline	Sub discipline	sub-discipline	Speciality		
01	FUNDAMENTAL BIOLOGY	BING	Biomedical Engineering	Engineering, biomedical		
				Medical informatics		
				Medical laboratory technology		
				Materials science, biomaterials		
				Microscopy		
		BIOG	Biochemistry/ Molecular & Cellular Biology & Genetics	Biochemical research methods		
				Biochemistry & molecular biology		
				Biophysics		
				Cell biology		
		DIVF	Misc. Fund. Biol.	Developmental biology		
				Anatomy & morphology		
				Genetics & heredity		
				Nutrition & dietetics		
		MIIN	Microbiology/ Virology/ Infections/ Bioprocesses	Physiology		
				Reproductive biology		
				Biotechnology & applied microbiology		
		NEUR	Neurosciences/ Psychology	Microbiology		
				Parasitology		
				Virology		
		02	MEDICAL RESEARCH	CANC	Oncology/ Hematology	Behavioral sciences
Neurosciences						
Psychology						
CHME	Medical Chemistry/ Pharmacy			Oncology		
				Hematology		
				Medicine, research & experimental		
ENDO	Endocrinology			Chemistry, medicinal		
				Medicine, legal		
				Pharmacology & pharmacy		
GMED	General Medicine/ Miscellaneous			Andrology		
				Endocrinology & metabolism		
				Medicine, general & internal		
IMMU	Immunology			Pathology		
				Radiology, nuclear medicine & medical imaging		
				Allergy		
INTE	Gastroenterology/ Cardiology/ Pneumology/ Surgery			Immunology		
				Anesthesiology		
				Cardiac & cardiovascular systems		
						Emergency medicine

			Gastroenterology
			Respiratory system
			Sport sciences
			Surgery
			Transplantation
			Urology & nephrology
			Peripheral vascular disease
	MSPE	Other Medical Specialties	Dentistry, oral surgery & medicine
			Dermatology & venereal diseases
			Infectious diseases
			Ophthalmology
			Orthopaedics
			Otorhinolaryngology
			Rehabilitation
			Rheumatology
			Tropical medicine
			Veterinary sciences
	SANT	Public Health/ Epidemiology/ Life Cycle/ Toxicology	Substance abuse
			Health care sciences & services
			Geriatrics & gerontology
			Public, environmental & occupational health
			Clinical neurology
			Obstetrics & gynecology
			Pediatrics
			Psychiatry
			Toxicology

ANNEX 2

construction of the sets - numerical example

This table illustrates the construction of the three sets (observed, ordered fields, random). The example here is purely fictitious, over-simplified and reduced to three levels of aggregation: journals (5 journals), specialities (2), sub-discipline (1). The example is based on a plausible distribution of citations. The three sets share the classification/structure including the size of journals and the scores of citations, but the original citation figure in the 'observed set' for one article is attributed to other articles: randomly for the 'random set', hierarchically along the arbitrary list of journals and further aggregates, in the 'ordered fields set'. For example the score 19 of the article #A6 in the original article goes to the article #A2 in the ordered set and #D7 in the random set. Relative rankings are based here on rank percentages with a random processing of ties - as recalled in the text, the results are sensitive to ranking options, since citation-type distributions yield a huge number of ties in the low frequency area.

In the last table, correlation is the usual correlation, analogous to Spearman since applied to rank percentages. Correlations amongst the three levels show, as expected, the intermediary situation of the 'observed set' between the random model (very high coefficients) and the hierarchical model (much lower). We can visually watch the instability of an 'excellence' subset (boldface), here arbitrarily fixed at the first decile, again intermediary between the built-in instability in the ordered set model, and the stability observed in the random model.

TABLE 2A CONSTRUCTION OF THE THREE SETS - FICTITIOUS EXAMPLE

CLASSIFICATION				OBSERVED SCI				benchmark: ORDERED FIELDS				benchmark: RANDOM SET			
SUBDISCIPL	SPECIALITY	JOURNAL	ARTICLE	citations	rk in subdisc	rk in speciality	rk in journal	citations	rk in subdisc	rk in speciality	rk in journal	citations	rk in subdisc	rk in speciality	rk in journal
X	1	A	A1	3	6,9	6,3	4,6	31	10,0	10,0	10,0	0	1,9	2,1	1,9
X	1	A	A2	1	4,1	2,6	1,0	19	9,8	9,5	9,1	4	7,8	6,8	7,3
X	1	A	A3	31	10,0	10,0	10,0	14	9,6	8,9	8,2	10	9,3	9,5	10,0
X	1	A	A4	10	9,3	8,4	7,3	12	9,4	8,4	7,3	1	5,0	4,7	4,6
X	1	A	A5	2	6,7	5,8	3,7	10	9,3	7,9	6,4	0	2,3	2,6	2,8
X	1	A	A6	19	9,8	9,5	9,1	8	9,1	7,3	5,5	2	5,6	5,2	5,5
X	1	A	A7	5	8,3	7,9	6,4	7	8,9	6,8	4,6	5	8,2	7,3	8,2
X	1	A	A8	2	6,5	5,2	2,8	6	8,7	6,3	3,7	0	1,5	1,5	1,0
X	1	A	A9	4	7,8	7,3	5,5	6	8,5	5,8	2,8	8	9,1	8,9	9,1
X	1	A	A10	12	9,4	8,9	8,2	5	8,3	5,2	1,9	2	6,1	5,8	6,4
X	1	A	A11	2	6,3	4,7	1,9	5	8,2	4,7	1,0	0	2,8	3,1	3,7
X	1	B	B1	1	3,9	2,1	4,0	4	8,0	4,2	10,0	0	1,4	1,0	1,0
X	1	B	B2	1	3,8	1,5	2,5	4	7,8	3,6	8,5	1	3,0	3,6	2,5
X	1	B	B3	2	6,1	4,2	8,5	4	7,6	3,1	7,0	12	9,4	10,0	10,0
X	1	B	B4	2	6,0	3,6	7,0	3	7,4	2,6	5,5	1	3,8	4,2	4,0
X	1	B	B5	4	7,6	6,8	10,0	3	7,2	2,1	4,0	7	8,9	8,4	8,5
X	1	B	B6	1	3,6	1,0	1,0	3	7,1	1,5	2,5	3	6,9	6,3	5,5
X	1	B	B7	2	5,8	3,1	5,5	3	6,9	1,0	1,0	6	8,7	7,9	7,0
X	2	C	C1	8	9,1	9,7	9,2	2	6,7	10,0	10,0	0	1,0	1,0	1,0
X	2	C	C2	1	3,4	4,8	2,6	2	6,5	9,7	9,2	4	8,0	8,5	8,4
X	2	C	C3	2	5,6	7,1	5,1	2	6,3	9,4	8,4	31	10,0	10,0	10,0
X	2	C	C4	2	5,4	6,8	4,3	2	6,1	9,1	7,5	1	4,7	4,8	5,9
X	2	C	C5	6	8,7	9,1	8,4	2	6,0	8,8	6,7	2	6,3	6,5	6,7
X	2	C	C6	14	9,6	10,0	10,0	2	5,8	8,5	5,9	1	4,3	4,2	4,3
X	2	C	C7	1	3,2	4,5	1,8	2	5,6	8,3	5,1	1	3,6	3,3	3,4
X	2	C	C8	3	7,4	8,0	7,5	2	5,4	8,0	4,3	0	1,2	1,3	1,8
X	2	C	C9	3	7,2	7,7	6,7	1	5,2	7,7	3,4	2	6,7	7,1	7,5
X	2	C	C10	1	3,0	4,2	1,0	1	5,0	7,4	2,6	1	4,5	4,5	5,1
X	2	C	C11	3	7,1	7,4	5,9	1	4,9	7,1	1,8	0	2,1	1,9	2,6
X	2	C	C12	1	5,2	6,5	3,4	1	4,7	6,8	1,0	5	8,3	8,8	9,2
X	2	D	D1	0	2,8	3,9	7,3	1	4,5	6,5	10,0	1	5,2	5,3	2,8
X	2	D	D2	0	2,6	3,6	6,4	1	4,3	6,2	9,1	2	5,4	5,6	3,7
X	2	D	D3	0	2,5	3,3	5,5	1	4,1	5,9	8,2	14	9,6	9,4	9,1
X	2	D	D4	0	2,3	3,0	4,6	1	3,9	5,6	7,3	2	6,5	6,8	6,4
X	2	D	D5	1	5,0	6,2	9,1	1	3,8	5,3	6,4	3	7,2	7,7	7,3
X	2	D	D6	5	8,2	8,5	10,0	1	3,6	5,1	5,5	2	6,0	6,2	5,5
X	2	D	D7	0	2,1	2,7	3,7	1	3,4	4,8	4,6	19	9,8	9,7	10,0
X	2	D	D8	0	1,9	2,4	2,8	1	3,2	4,5	3,7	0	2,6	2,4	1,0
X	2	D	D9	0	1,7	2,2	1,9	1	3,0	4,2	2,8	2	5,8	5,9	4,6
X	2	D	D10	1	4,9	5,9	8,2	0	2,8	3,9	1,9	6	8,5	9,1	8,2
X	2	D	D11	0	1,5	1,9	1,0	0	2,6	3,6	1,0	1	4,1	3,9	1,9
X	2	E	E1	1	4,7	5,6	6,6	0	2,5	3,3	10,0	3	7,1	7,4	7,8
X	2	E	E2	0	1,4	1,6	3,3	0	2,3	3,0	8,9	1	3,4	3,0	4,4
X	2	E	E3	4	8,0	8,3	7,8	0	2,1	2,7	7,8	1	4,7	4,8	6,6
X	2	E	E4	1	4,5	5,3	5,5	0	1,9	2,4	6,6	3	7,4	8,0	8,9
X	2	E	E5	7	8,9	9,4	10,0	0	1,7	2,2	5,5	1	3,9	3,6	5,5
X	2	E	E6	0	1,2	1,3	2,1	0	1,5	1,9	4,4	0	1,7	1,6	1,0
X	2	E	E7	0	1,0	1,0	1,0	0	1,4	1,6	3,3	0	2,5	2,2	2,1
X	2	E	E8	6	8,5	8,8	8,9	0	1,2	1,3	2,1	4	7,6	8,3	10,0
X	2	E	E9	1	4,3	5,1	4,4	0	1,0	1,0	1,0	1	3,2	2,7	3,3

TABLE 2B CORRELATIONS OF RANKINGS AMONGST LEVELS

option for ties:	corr	spec	jnal	corr	spec	jnal	corr	spec	jnal
random		0,91	0,74	0,54	0,20		0,99	0,94	
	spec		0,77	spec	0,40		spec		0,95