B04.010

Handbook of Quantitative Science and Technology Research

The Use of Publication and Patent Statistics in Studies of S&T Systems

edited by

Henk F. Moed

Centre for Science and Technology Studies,
University of Leiden, The Netherlands

Wolfgang Glänzel

Steunpunt O&O Statistieken,
K.U. Leuven, Belgium

and

Ulrich Schmoch

Fraunhofer Institute for Systems and Innovation Research,
Karlsruhe, Germany

KLUWER ACADEMIC PUBLISHERS
DORDRECHT / BOSTON / LONDON

Library of Congress Cataloging-in-Publication Data

1. Van Raan
2. Grupp/Mogee

ISBN 1-4020-2702-8 (HB)
ISBN 1-4020-2755-9 (e-book)

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Sold and distributed in North, Central and South America
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

All Rights Reserved
© 2004 Kluwer Academic Publishers
No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed in the Netherlands.

Chapter 1

MEASURING SCIENCE

Capita Selecta of Current Main Issues

Anthony F.J. van Raan

Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands
E-mail: vanraan@cwts.leidenuniv.nl

Abstract: After a review of developments in the quantitative study of science, particularly since the early 1970s, I focus on two current main lines of 'measuring science' based on bibliometric analysis. With the developments in the Leiden group as an example of daily practice, the measurement of research performance and, particularly, the importance of indicator standardisation are discussed, including aspects such as interdisciplinary relations, collaboration, 'knowledge users'. Several important problems are addressed: language bias; timeliness; comparability of different research systems; statistical issues; and the 'theory–invariance' of indicators. Next, an introduction to the mapping of scientific fields is presented. Here basic concepts and issues of practical application of these 'science maps' are addressed. This contribution is concluded with general observations on current and near-future developments, including network–based approaches, necessary 'next steps' are formulated, and an answer is given to the question 'Can science be measured?'

1. TOWARD A METRIC OF SCIENCE REVISITED[1]

From the early sixties onwards we see a strong increase in quantitative material on the state-of-the art in science and technology. National institutes of statistics, UNESCO, OECD, and the European Commission are main examples of organisations starting to collect systematically data on the

[1] The book *Toward a metric of Science: The Advent of Science Indicators* (Elkana et al., 1978) has always been a one of my major sources of inspiration. This contribution to the Handbook is based on earlier publications by the author (Van Raan 2000a; Van Raan and Noyons 2002).

development of science and technology. An important milestone is the first issue of the OECD 'Frascati Manual' (OECD, 1963), a handbook devoted to the development of a standard practice for surveys of the measurement of scientific and technical activities. At the same time, and strongly related to this data explosion, the quantitative appraisal of current science gains influence. As a genre in the study of the history of science, the quantitative approach of the development of science, 'scientometrics', is certainly not new. A remarkable early piece of work is "Histoire des sciences et des savants depuis deux siècles". The author, Alphonse de Candolle (1873), described the changes in the scientific strength of nations by membership of scientific societies, and he tried to find 'environmental factors' of all kinds (even including the role of the celibate) for the scientific success of a nation. Later, in the 1920s, Lotka (1929) published his famous work on the productivity of chemistry researchers. Here scientometrics is clearly differentiated into 'bibliometrics'.

Undoubtedly the invention of the *Science Citation Index* by Eugene Garfield is a major breakthrough (Wouters, 1999). This invention enabled statistical analyses of the scientific literature on a very large scale. It marks the rise of bibliometrics as a powerful field within the studies of science. Such great scientists as Derek de Solla Price and Robert Merton recognised the value of Garfield's invention, Price from the perspective of contemporaneous history of science, Merton from the perspective of normative sociology.

Scientists are fascinated by basic features such as simplicity, symmetry, harmony, and order. The *Science Citation Index* enabled De Solla Price to start with the development of a 'physical approach' to science, in which he tried to find laws to predict further developments, inspired by the ideas of Newtonian and statistical mechanics. In this perspective, quantitative measures of science, '*indicators*', are guides to find and, as a crucial next step, to understand such basic features. The most basic feature concerns the cognitive dimension: the development of content and structure of science. More on the mundane surface science indicators relate to the social dimension of science, in particular to aspects formulated in questions such as 'How many researchers? How much money is spent on science? How 'good' are research groups? How does communication in science work, particularly what is the role of books, journals, conferences (Borgman, 1990)? And longer than we often realise there is another question: 'What is the economic profit of scientific activities?' A landmark in the development of science indicators is the first publication in a biennial series of the *Science Indicators* Report in 1973. Stimulated by the success achieved by economists in developing quantitative measures of political significance (e.g., unemployment, GNP), the US National Science Board started this indicator

report series in which we find more emphasis on the demographic and economic state of science than on the cognitive state of science (National Science Board, 1973).

Making quantitative indicators of anything thinkable fascinates some people and horrifies others as being nonsense and taking us back to the cabbalistic magic number world of Paracelsus. But there are famous classical pronouncements to support the attempt to measure things. Horace (65–5 BC): "There is a measure in all things" (Est modus in rebus), Johannes Kepler (1597): "The mind comprehends a thing the more correctly the closer the thing approaches toward pure quantity as its origin", and, from the place where I live and work, Leiden, the discoverer of superconductivity, Heike Kamerlingh Onnes (1882): "Measuring is knowing".

There is no final theory of science providing *the* methodology of measurement. It is a returning hype in the social studies of science to incite the scientific community with this observation. But are we really troubled by this poverty of theoretical content? I don't think so (van Raan, 1997). Do not expect a classical mechanics of scientometrics. With very high probability: it does not exist. The absence of any explicit theory to guide the making and use of indicators may not be good, but the adoption of a single one, for instance, a trendy dominating 'theory', is likely to be worse (Holton, 1978). It is normal practice in empirical science to begin a search without a theoretical clarification and try to establish a model to explain the findings later. Certainly in such measurements we do have at least implicit basic ideas about 'how things work' and the same is true for the construction and use of science indicators. Therefore it is crucial to make these implicit assumptions clear to the outside world. This will allow us to turn the absence of a general theory of the development of science into a very profitable situation, in the words of Gerald Holton: 'perhaps indicators may be developed eventually that are *invariant* with respect to theoretical models. They and only they allow rival theories to be put to empirical tests'. To put it more bluntly: we cannot develop a sound theoretical model of the 'sociology of knowledge' yet, as we simply need more empirical work based on the richness of available and future data in order to develop a better quantitative understanding of the processes by which science and society mutually influence each other's progress. In this contribution I will argue that advanced bibliometric indicators approach the above characteristic of invariance.

What is the difference between data and indicators? An indicator is the result of a specific mathematical operation (often simple arithmetic) with data. The mere number of citations of one publication in a certain time period is *data*. The measure in which such citation counts of all publications of a research group in a particular field are normalised to citation counts of

all publications worldwide in the same field, is an *indicator*. An indicator is a measure that explicitly addresses some assumption. In our example the assumption is: this is the way to calculate the international scientific influence of a research group. So, to begin with, we need to answer the question: what features of science can be given a numerical expression? Thus indicators can not exist without a specific goal in mind, they have to address specific questions, and thus they have to be created to gauge important 'forces'; for example, how scientific progress is related to specific cognitive as well as socio-economic aspects. Indicators must be problem driven, otherwise they are useless. They have to describe the recent past in such a way that they can guide us, can inform us about the near future. A second and more fundamental role of indicators is their possibility to test aspects of theories and models of scientific development and its interaction with society. In this sense, indicators are not only tools for science policy makers and research managers, but also instruments in the study of science. But we also have to realise that science indicators do not answer typical epistemological questions such as: How do scientists decide what will be called a scientific fact? How do scientists decide whether a particular observation supports or contradicts a theory? How do scientists come to accept certain methods or scientific instruments as valid means of attaining knowledge? How does knowledge selectively accumulate? (Cole et al., 1978).

De Solla Price (1978) strikingly described the mission of the indicator maker: find the most simple pattern in the data at hand, and then look for the more complex patterns which modify the first. What should be constructed from the data is not a number but a pattern, a cluster of points on a map, a peak on a graph, a correlation of significant elements on a matrix, a qualitative similarity between two histograms. If these patterns are found the next step is to suggest models that produce such patterns and to test these models by further data. A numerical indicator or an indicative pattern, standing alone, has little significance. The data must be given perspective: the change of an indicator with time, or different rates of change of two different indicators. Crucial is that numerical quantities are replaced by geometrical or topological objects or relations (Ziman, 1978).

We know already from the early indicator work that these 'simple patterns' exist: the rank of countries by the number of publications is remarkably stable from year to year (Braun et al., 1995). The absolute size of the scientific research activity in the number of publications of any nation is in very good agreement with its electrical power consumption in kilowatt-hours, indicating that scientific power, economic power, and national wealth are strongly related.

More or less at the same time as the above thoughts on the metric of science, Francis Narin coined the concept of 'evaluative bibliometrics'. His pioneering work on the development of research performance indicators (Narin, 1976, 1978), mainly on the macro level, i.e., the performance of countries, was a new, important breakthrough which contributed substantially to the measurement of scientific activities. In 1978 Tibor Braun founded the journal *Scientometrics*. This event marks the emancipation of the field of quantitative studies of science. Also in journals such as *Research Policy* and the *Journal of the American Society for Information Science* we find more and more publications about 'measuring science', and most of them are on topics that are still very relevant. We mention, without being exhaustive, the seminal papers in the 1970s on the development of 'relational' methods such as co-citation analysis for the mapping of scientific fields (Small, 1973), on scientific collaboration by deB. Beaver and colleagues (Beaver, 1978), on measuring the growth of science (Moravcsik, 1975; Gilbert, 1978), the meaning of citation patterns for assessing scientific progress (Moravcsik and Murugesan, 1978), and on mobility in science (Vláchy, 1979).

In the early eighties we see the rapid rise of co-citation analysis (Small and Greenlee, 1980; Sullivan et al., 1980; Price, 1981; White and Griffith, 1981; Noma, 1982; McCain, 1984) and of co-word analysis (Callon et al., 1983; Rip and Courtial, 1984), an increasing emphasis on advanced statistical analysis of scientometric parameters (Haitun, 1982; Schubert and Glänzel, 1983), the application of bibliometric methods in the social sciences (Peritz, 1983), indicators of interdisciplinary research (Porter and Chubin, 1985), and comparison of peer opinions and bibliometric indicators (Koenig, 1983).

An important further breakthrough was the work of Martin and Irvine (1983) on the application of science indicators at the level of research groups. Around the same time (the beginning of the eighties) our Leiden institute had also started with bibliometric analysis oriented on research groups (Moed et al., 1983) and Braun and co-workers focused on the scientific strength of countries in a wide range of research fields (Braun et al., 1988).

Now, almost thirty years after Narin's *Evaluative Bibliometrics*, twenty-five years after the publication of *Toward a Metric of Science: The Advent of Science Indicators* (Elkana et al., 1978), twenty years after Martin and Irvine (1983), and fifteen years after the *Handbook of Quantitative Studies of Science and Technology* (van Raan, 1988) we may state *plus ça change, plus c'est la même chose*. What changed is the very significant progress in application oriented indicator work based on the enormous increase of available data and, above all, the almost unbelievable, compared with the

situation in the seventies, increase of computing power and electronic facilities. I hope this contribution and handbook as a whole will prove this progress convincingly.

What also changed is the method of publishing. Electronic publishing and electronic archives mark an area of new information technology. I expect that most changes will be primarily technological but not conceptual. Publication via journals of high reputation is in most fields of science crucial for receiving professional recognition. That will remain so in the rapidly developing electronic area. A much more revolutionary change in science is the increasing availability and sharing of research results and, particularly, research data.

What remained, however, are some of the most fundamental questions. For instance: do science maps derived from citation and/or concept–similarity data have reality in a strictly spatial sense? In other words, do measures of similarity imply the existence of a metric space? This question brings us to an even more fundamental problem: the ontological status of maps of science will remain speculative until more has been learned about the structure of the brain itself (de Solla Price, 1978). For instance, it remains fascinating that science can be represented quite well in 2D space. Why is that so? Because our own brain is a (folded) two dimensional structure?

And yes, some old wishes have come true. It is now possible to make a time series of science maps, a 'science cinematography' that enables us to examine shifts in clusters over time and to investigate the nature of change of research themes and specialties. Short term extrapolation may be feasible.

A new development is a 'physical' network approach to analysing publication and citation relations. Recently we reported some first results on network characteristics of a reference based, bibliographically coupled publication network structure (van Raan, 2003). It was found that this network of clustered publications shows different topologies depending on the age of the references used for building the network. Also progress is made in the understanding of the statistics of citation distributions. This is of crucial importance, as it is directly related to the 'wiring' (citations) of the 'nodes' (publications) in the network structure of science. A two-step competition process is applied as a model for explaining the distribution of citations ('income') over publications ('work'). A distribution function of citing publications is found which corresponds very well to the empirical data. It is not a power law, but a modified Bessel function. This model has a more generic value, particularly in economics for explaining observed income distributions (van Raan, 2001).

In this contribution we focus on two main lines of 'measuring science' based on bibliometric analysis. First, in the next section, we discuss the

measurement of research performance, including aspects such as interdisciplinarity, collaboration, 'knowledge users'. I address several important problems: language bias; timeliness; comparability of different research system; statistical issues; the relation between bibliometric finding and peer judgements. The latter issue is followed by a first discussion of Holton's ideal of 'theory invariant' indicators. In Section 3 an introduction to the mapping of scientific fields is presented. I discuss basic concepts and issues of the practical application of these 'science maps'. Finally, in Section 4 this contribution is concluded with some general observations on current and near-future developments, particularly in relation to network–based approaches and growth phenomena. Necessary 'next steps' are formulated. But first, back to the basics.

2. BIBLIOMETRIC MEASUREMENT OF SCIENTIFIC PERFORMANCE

2.1 Basic Concepts

The rationale of the bibliometric approach to measuring scientific performance presented in this contribution is as follows. Scientific progress can be defined as the substantial increase of our knowledge about 'everything'. In broad outline we discern basic knowledge ('understanding') and applicable knowledge ('use'). This knowledge can be tacit ('craftsmanship') or codified ('archived & publicly accessible'). Scientists have communicated (and codified) their findings in a relatively orderly, well defined way since the 17th century. Particularly is the phenomenon of serial literature crucial: publications in international journals. Thus communication, i.e., exchange of research results, is a crucial aspect of the scientific endeavour. Publications are not the only, but certainly very important elements, in this process of knowledge exchange.

Each year about 1,000,000 publications are added to the scientific archive of this planet. This number and also numbers for sub-sets of science (fields, institutes) are in many cases sufficiently high to allow quantitative analyses yielding statistically significant findings. Publications offer usable elements for 'measuring' important aspects of science: author names, institutional addresses, journal (which indicates not only the field of research but also status!), references (citations), concepts (keywords, keyword combinations). Although not perfect, we adopt a publication as a 'building block' of science and as a source of data. This approach clearly defines the basic assumptions of bibliometrics (Kostoff, 1995). Thus bibliometric

assessment of research performance is based on one central assumption: scientists who have to say something important do publish their findings vigorously in the open international journal ('serial') literature. This choice introduces unavoidably a 'bibliometrically limited view of a complex reality'. For instance, journal articles are not in all fields the main carrier of scientific knowledge; they are not 'equivalent' elements in the scientific process, they differ widely in importance; and they are challenged as the 'gold standard' by new types of publication behaviour, particularly electronic publishing. However, the daily practice of scientific research shows that inspired scientists in most cases, and particularly in the natural sciences and medical research fields, go for publication in the better and, if possible, the best journals. A similar situation is developing in the social and behavioural sciences (Glänzel, 1996; Hicks, 1999), engineering and, to a lesser extent, in the humanities. This observation is confirmed by many years of experience in peer review based research evaluation procedures.

Work of at least some importance provokes reactions of colleagues. They are the international forum, the 'invisible college', by which research results are discussed. Often these colleagues play their role as a member of the invisible college by referring in their own work to earlier work of other scientists. This process of citation is a complex one, and it certainly does not provide an 'ideal' monitor on scientific performance (MacRoberts and MacRoberts, 1996). This is particularly the case at a statistically low aggregation level, e.g., the individual researcher. But the application of citation analysis to the work, the 'oeuvre' of *a group of researchers as a whole over a longer period of time*, does yield in many situations a strong indicator of scientific performance.

Citation analysis is based on reference practices of scientists. The motives for giving (or not giving) a reference to a particular article may vary considerably (Brooks, 1986; MacRoberts and MacRoberts, 1988; Vinkler, 1998). There is, however, sufficient evidence that these 'reference motives' are not so different or 'randomly given' to such an extent that the phenomenon of citation would lose its role as a reliable measure of impact (van Raan, 1998).

Why bibliometric analysis of research performance? Peer review undoubtedly is and has to remain the principal procedure of quality judgment. But peer review and related expert–based judgments may have serious shortcomings and disadvantages (Moxham and Anderson, 1992; Horrobin, 1990). Subjectivity, i.e., dependence of the outcomes on the choice of individual committee members, is one of the major problems. This dependence may result in conflicts of interests, unawareness of quality, or a negative bias against younger people or newcomers to the field. Basically, the methodological problem of determining the quality of a subject is still far

from solved, as illustrated by the results of re-review of previously granted research proposals, see, for instance, Nederhof (1988). I do not plead for a replacement of peer review by bibliometric analysis. Subjective aspects are not merely negative. In any judgment there must be room for the intuitive insights of experts. I claim, however, that for a substantial improvement of decision making an advanced bibliometric method, such as presented in this contribution has to be used in parallel with a peer–based evaluation procedure.

The earlier mentioned pioneering work of Narin (1976) and of Martin and Irvin (1983) clearly showed that the most crucial parameter in the assessment of research performance is international scientific influence. Citation–based bibliometric analysis provides indicators of international impact, influence. This can be regarded as, at least, one crucial aspect of scientific quality, and thus a 'proxy' of quality as follows from a long standing experience in bibliometric analysis. Perhaps this is the best answer of the classical question posed by Eugene Garfield (1979): 'Is citation analysis a legitimate evaluation tool?' Therefore we have developed standardised bibliometric procedures for assessing research performance within the framework of international influence. Undoubtedly, this approach does not provide us an ideal instrument, working perfectly in all fields under all circumstances. But the approach presented in this contribution works very well in the large majority of the natural, the medical, the applied, and the behavioural sciences. These fields of science are the most cost intensive and the ones with the strongest socio-economic impact. For a recent application of bibliometric research performance assessment in a typical applied field such as food and nutrition research we refer to Van Raan and Van Leeuwen (2002). The application of bibliometric analysis in the humanities is discussed by Moed et al. (2002).

A first and good indication of whether bibliometric analysis is applicable to a specific field is provided by the publication characteristics of the field; in particular, the role of *international* refereed journals. If international journals are a dominating or at least a major means of communication in a field, then in most cases bibliometric analysis is applicable. Therefore it is important to study the 'publication practices' of a research group, department, or institute, in order to establish whether bibliometric analysis can be applied. A practical measure here is the share of CI-covered[2]

[2] The Science Citation Index, the Social Science Citation Index, the Arts & Humanities Citation Index, and the 'specialty' citation indexes (CompuMath, Biochemistry and Biophysics, Biotechnology, Chemistry, Material Science, Neurosciences) are produced and published by the Institute for Scientific Information (ISI/Thomson Scientific) in

publications in the total research output. For 'not-CI covered publications' a restricted type of analysis is possible, in so far as these publications are cited by articles in journals covered by the CI.

We have already noticed that journal publications are challenged as the 'gold standard' in science as the worldwide web has changed scientific communication. Researchers use the web for information seeking, and in addition to the above mentioned 'not-CI covered publications' there is an enormous number of further publications and data included in institutional and personal websites. Thus next to citation analysis, in the use of data provided via the internet, 'webometrics' offers interesting additional opportunities to aid citation–based bibliometric analysis in evaluation and mapping approaches (Björneborn and Ingwersen, 2001; Bar-Ilan, 2001; Thelwall and Smith, 2002; Thelwall and Harries, 2003).

The Leiden group has gained an extensive experience in bibliometric analysis. In a period of almost 20 years we have studied the research performance of many thousands of research groups, worldwide. By all these activities an empirical gold mine was created. We first discuss our methodology in the next section, and in Section 2.3.5 we explain why we think that this methodology has yielded indicators which, at least, approach Holton's ideal of theory–invariant measures.

2.2 Details of the Methodology

One of the most crucial objectives in bibliometric analysis is to arrive at a consistent and *standardised set* of indicators. The methodology presented in this section is driven by this motive. Research output is defined as the number of articles of the institute, as far as covered by the *Science Citation Index* (SCI) and all its related databases (see footnote 3). As 'article' we consider the following publication types: normal articles (including proceedings papers published in journals); letters; notes; and reviews (but not meeting abstracts, obituaries, corrections, editorials, etc.).

I take the results of a recent analysis by our institute of a German medical research institute as an example (over the period 1992–2000). Table 1.1 shows the number of papers published, ***P***, which is also a first indication of the size of an institute. This number is about 250 per year. Next we find the total number of citations, ***C***, received by ***P*** in the indicated period, *and corrected for self-citations*. For papers published in 1996 citations are counted during the period 1996–2000, for 1997 papers citations in 1997–

Philadelphia. Throughout this paper we use the term 'CI' (Citation Index) for the above set of databases.

2000, and so on. For the outsider this looks like 'just counting numbers'. But the reliable establishment of even these two basic indicators is far from trivial. Verification is crucial in order to remove errors and to detect incompleteness of addresses of research organisations, departments, groups.

In citation analysis an entire range of pitfalls and sources of error is lurking. We refer to Van Raan (1996) for the many methodological and technical problems which have to be solved in order to conduct a bibliometric analysis properly. There is ample empirical evidence that in the natural and life sciences, basic as well as applied, the average 'peak' in the number of citations is in the third or fourth year after publication. Therefore a five-year period is appropriate for impact assessment. A trend analysis is then based on 'moving' and partially overlapping five-year periods, as presented in Table 1.1.

The third and fourth indicators are the average number of citations per publication ***(CPP)***, again without self-citations, and the percentage of not-cited papers, ***% Pnc***. We stress that this percentage of non-cited papers concerns, like all other indicators, the given time period. It is possible that publications not cited within such a time period will be cited after a longer time. This is clearly visible when comparing this indicator for the five-year periods (e.g., 1996–2000: 30%) with that of the whole period (1992–2000: 21%). The values found for this medical research institute are quite normal.

How do we know that a certain number of citations, or a certain value of citations-per-publication is low or high? To answer this question we have to make a comparison with (or normalisation to) a well chosen international reference value, and thus to establish a reliable measure of *relative, internationally field–normalised impact*. Another reason for normalising the measured impact of an institute (***CPP***) to international reference values is that overall worldwide citation rates are increasing. I stress, however, that the distribution of citations over publications is skew and therefore we have to be careful with the use of mean values. In Section 2.3 a short discussion of statistical problems in bibliometric analysis is given.

First, the average citation rate of all papers (worldwide) in the journals in which the institute has published (***JCSm***, the **m**ean **J**ournal **C**itation **S**core of the institute's 'journal set', and ***JCS*** for one specific journal) is calculated. Thus this indicator ***JCSm*** defines a worldwide reference level for the citation rate of the institute. It is calculated in the same way as ***CPP***, but now for all publications in a set of journals (see van Raan, 1996, 2003). A novel and unique aspect is that we take into account the type of paper (e.g., letters, normal article, review) *as well as* the specific years in which the papers were published. This is necessary, because the average impact of journals may have considerable annual fluctuations and large differences per article type, see Moed and Van Leeuwen (1995, 1996).

With help of the ratio ***CPP/JCSm*** we observe whether the measured impact is *above* or *below* the international average. However, comparison of the institute's citation rate ***(CPP)*** with the average citation rate of its journal set (***JCSm***) introduces a specific problem related to journal status (Lewison, 2002). For instance, if a research group publishes in prestigious (high impact) journals, and another group in rather mediocre journals, the citation rate of articles published by both groups may be equal *relative to* the average citation rate of their respective journal sets. But generally one would argue that the first group evidently performs better than the second. Therefore we developed a second international reference level, a *field*–based world average ***FCS***, and ***FCSm*** in the case in which more fields are involved. This indicator is based on the citation rate of *all* papers (worldwide) published in *all* journals of the field(s)[3] in which the institute is active, and not only the journals in which the institute's researchers publish their papers. Thus, for a publication in a less prestigious journal one may have a (relatively) high ***CPP/JCSm*** but a lower ***CPP/FCSm***, and for a publication in a more prestigious journal one may expect a higher ***CPP/FCSm*** because publications in a prestigious journal will generally have an impact above the field–specific average.

Table 1.1. Bibliometric analysis of a medical research institute, 1992–2000

Period	*P*	*C*	*CPP*	*%P nc*	*CPP/ JCSm*	*CPP/ FCSm*	*CPP/ D-FCSm*	*JCSm/ FCSm*	*% sc*
1992 – 00	2,245	43,665	19.45	21	1.26	1.95	1.85	1.55	18
1992 – 96	1,080	11,151	10.33	36	1.27	2.02	1.95	1.58	22
1993 – 97	1,198	12,794	10.68	34	1.24	2.03	1.92	1.63	21
1994 – 98	1,261	12,217	9.69	32	1.19	1.85	1.72	1.55	22
1995 – 99	1,350	13,709	10.15	31	1.21	1.89	1.76	1.56	21
1996 – 00	1,410	14,815	10.51	30	1.20	1.91	1.76	1.59	21

[3] We use here the definition of fields based on a classification of scientific journals into *categories* developed by ISI. Although this classification is not perfect, it provides a clear and 'fixed' consistent field definition suitable for automated procedures within our data system. A more 'real world', user oriented, definition of fields can be provided by the bibliometric mapping methodology discussed in Section 3 of this contribution.

The same procedure is used as applied in the calculation of ***JCSm***. Often an institute is active in more than one field. In such cases a weighted average value is calculated, the weights being determined by the total number of papers published by the institute in each field. For instance, if the institute publishes in journals belonging to genetics as well as to cell biology, then the ***FCSm*** of this institute will be based on both field averages. Thus the indicator ***FCSm*** represents a *world average*[4] in a specific (combination of) field(s). It is also possible to calculate ***FCSm*** for a specific country or for the European Union. The example discussed in this paper concerns a German medical research institute, and for this institute we calculated the Germany–specific ***FCSm*** value, ***D-FCSm***.

As in the case of ***CPP/JCSm***, if the ratio ***CPP/FCSm*** is above 1.0 the impact of the institute's papers exceeds the field–based (i.e., *all* journals in the field) world average. We observe in Table 1.1 that the ***CPP/JCSm*** is 1.20, ***CPP/FCSm*** 1.91 and ***CPP/D-FCSm*** is 1.76 in the last period 1996–2000. These results show that the institute is performing well above international average. The ratio ***JCSm/FCSm*** is also an interesting indicator. If it is above 1.0, the mean citation score of the institute's journal set exceeds the mean citation score of all papers published in the field(s) to which the journals belong. For the institute this ratio is around 1.59. This means that the institute publishes in journals with, generally, a high impact. The last indicator shows the percentages of self-citations (***%Sc***). About thirty percent is normal, so the self-citation rates for this institute are certainly not high (about 20%).

A general, and important, observation is the 'stability' over time of most indicators. This is quite typical, particularly for groups and institutes of high reputation. The conclusion to be drawn from this observation is that the indicators are not a 'noisy set of measures' but apparently represent an enduring characteristic of scientific work, including communication practices.

I regard the *internationally standardised impact indicator* ***CPP/FCSm*** as our 'crown' indicator. This indicator enables us to observe immediately whether the performance of a research group or institute is significantly far below (indicator value < 0.5), below (indicator value between 0.5 and 0.8), about (between 0.8 and 1.2), above (between 1.2 and 1.5), or far above (>1.5) the international impact standard of the field. I stress, however, that

[4] About 80 percent of all CI-covered papers is authored by scientists from the United States, Western Europe, Japan, Canada, and Australia. Therefore our 'world average' is dominated by the Western world.

for the interpretation of the measured impact value one has to take into account the *aggregation level of the entity* under study. The higher the aggregation level the larger the volume of publications and the more difficult it is to have an impact significantly above the international level. Based on our long standing experiences, I can say the following. At the 'meso level' (e.g., a university, faculty, or large institute, with about 500 or more publications per year), a ***CPP/FCSm*** value above 1.2 means that the institute's impact as a whole is significantly above the (western) world average. With a ***CPP/FCSm*** value above 1.5, such as in our example, the institute can be considered to be scientifically strong, with a high probability of finding very good to excellent groups. Thus the next step in a research performance analysis is a breakdown of the institution into smaller units, i.e., research groups. Therefore the bibliometric analysis has to be applied on the basis of institutional input data about personnel and composition of groups. The algorithms then can be repeated on the lowest but most important aggregation level, the research group. In most cases the volume of publications at this level is 10 to 20 per year.

Particularly at this lower aggregation level the verification of the data is crucial (e.g., correct assignment of publications to research groups, completeness of publications sets). In our institute we have developed standardised procedures for carrying out the analysis as conscientiously as possible. These procedures are discussed thoroughly beforehand with the client institutes.

At the group level a ***CPP/FCSm*** value above 2 indicates a very strong group, and above 3 the groups can be, generally, considered to be excellent and comparable to the top groups at the best US universities. If the threshold value for the ***CPP/FCSm*** indicator is set at 3.0, excellent groups can be identified with high probability (van Raan, 2000a). As an additional indicator of *scientific excellence* the number of publications within the *top 10%* of the worldwide impact distribution of the field concerned is determined for the target entity (see Noyons et al., 2003). In the calculation of this indicator the entire citation distribution function is taken into account, thus providing a better statistical measure than those based on mean values (see Section 2.3).

Science is, for a major part, teamwork. Particularly is international collaboration essential, not only for the working floor but also as policy for countries to keep pace in scientific progress (Vinkler, 1993; Arunachalam et al., 1994; Melin and Persson, 1996; Glänzel, 2001). For all the above indicators we also perform a breakdown into types of *scientific co-operation* according to the publication addresses: work by only the unit itself; in a national collaboration; or in an international collaboration. Generally one observes the highest impact for publications in international collaboration.

A further important step is the *breakdown* of the institute's output *into* research fields. This provides a clear impression of the research scope or 'profile' of the institute. Such a *spectral analysis* of the output is based on the simple fact of an the institute's researchers publishing in journals of many different fields. Our example, the German medical research institute, is a centre for molecular research oriented towards medicine. The researchers of this institute are working in a typical interdisciplinary environment. The institute's publications are published in a wide range of fields: biochemistry and molecular biology, genetics and heredity, oncology, cell biology, and so on. By ranking fields according to their size (in terms of numbers of publications) in a graphical display, we construct the research profile of the institute. Furthermore, we provide the field–normalised impact values of the institute's research in these different fields with help of ***CPP/FCSm***.

Figure 1.1 shows the results of this *bibliometric spectroscopy*. Thus it becomes immediately visible in which fields within its interdisciplinary research profile the institute has a high (or lower) performance. We observe the scientific strength of the target institute: its performance in the top four fields is high to very high. If we find a smaller field with a relatively low impact (i.e., a field in the lower part, the 'tail' of the profile), this does not necessarily mean that the (few) publications of the institute in this particular field are 'bad'. Often these small fields in a profile are those that are quite 'remote' from the institute's core fields. They are, so to say, peripheral fields. In such a case the institute's researchers may not belong to the dominating international research community of those fields, and their work may be not be cited as frequently as the work of these dominating ('card holding') community members.

In a similar way a breakdown of *the citing publications* into fields of science is made, which yields a profile of the *users* of scientific results (as far as represented by citing publications). This 'knowledge users' profile is a powerful indicator of *who* is using *which* research results, *where* (in which fields) and *when*. Thus it analyses *knowledge diffusion and knowledge use* and it indicates further interdisciplinary 'bridges', potential collaboration, and possible 'markets' in the case of applied research. For an example of these 'knowledge user profiles' I refer to Van Raan and Van Leeuwen (2002). The construction of these profiles can be considered also as an empirical method of studying interdisciplinary aspects of research. For instance, the distribution of the lengths of the field–specific bars in the profile can be used as a measure of interdisciplinarity.

Figure 1.1. Research profile of a medical research institute, 1992–2000

2.3 Important Issues in Applications...and What About Theory?

2.3.1 Language bias

Recent work (Grupp et al., 2001; van Leeuwen et al., 2001) shows that the utmost care must be taken in interpreting bibliometric data in a

comparative evaluation of national research systems (May, 1997). The measured value of impact indicators of research activities at the level of an institution and even of a country strongly depends upon whether one includes or excludes publications in CI-covered journals written in languages other than English. This is owed to the simple fact of the CI covering non-English language journals of which the papers have a considerably lower impact than those in the English language journals. Differences of measured impact of the order of 10 to 20% are possible. These findings clearly illustrate that indicators, even at the 'macro level', need to be interpreted against the background of their inherent limitations, such as, in this case, the effects of the language of publication.

2.3.2 Timeliness of the analysis

A frequently posed question concerns the 'delay problem': Does bibliometric analysis suffer from a substantial 'delay' in the measurement of research performance (Egghe and Rousseau, 2000)? An answer to this question first needs a further refinement: delay compared to what? To the average 'processing time' of a publication? To the average 'running time' of a project? Or to peer review 'time cycles'? The entire process starting with scientific activities and leading to 'publishable' results, the writing of an article, the submission of the article, the publication of the article, the citations to the article, varies considerably for the different fields of science, and often within a field. Depending on type of activities and type of results it may take years. But during that time the work is improved, the whole process time can not be regarded is a 'delay' or a 'waste of time'. Furthermore, the average duration of a major research project is about 4 years, and the same is the case for most peer review time cycles. Also, during the publication process the awareness of scientific community (and peers!) evolves (e.g., average time between field–specific conferences etc.). We also have cases where the analysis can be performed almost in 'real time', as illustrated by an example[5] of a recent physics paper with citing articles published in the same year as the cited publication.

[5] Publication in *Physical Review Letters*, vol. 88, page 138701, year of publication 2002. The first citing articles are in the same year as the cited publication, we show the first four:
Marc Barthélemy *et al.*, Phys. Rev. E **66**, 056110 (2002);
Petter Holme, Phys. Rev. E **66**, 036119 (2002);
Holger Ebel *et al.*, Phys. Rev. E **66**, 035103 (2002;
Haijun Zhou, Phys. Rev. E **66**, 016125 (2002).

The above implies that 'bibliometric awareness' does not necessarily take more time than 'peer awareness'. Moreover, the bibliometric system itself proves empirically the robustness of the method simply by showing that in many cases indicators, based on citation analysis, for universities, institutes, and larger research groups, are remarkably stable, as illustrated by the results presented in Table 1.1. We conclude that recent past performance is a reliable predictor for near future performance.

We also have to keep in mind that the importance of a publication does not necessarily appear immediately, even to peers, and that identification of quality may take considerable time (Garfield, 1980). An interesting phenomenon in this respect is the 'Sleeping Beauty in Science', a publication that goes unnoticed ('sleeps') for a long time and then, almost suddenly, attracts a lot of attention ('is awakened by the prince'). Recently the first extensive measurements of 'delayed recognition papers' (Glänzel et al., 2003) and the occurrence of Sleeping Beauties in the science literature (van Raan, 2004) have been reported. In the latter work an 'awakening' probability function is derived from the measurements, and the 'most extreme Sleeping Beauty up to now' identified.

2.3.3 Comparability of the different research systems

It is often quite problematic to understand and 'unravel' the structure of a research organisation in terms of 'real' units such as departments or research groups. There are major differences in research systems between countries. For instance, the University of London is no longer a university in the usual sense. It is an 'umbrella organisation' covering several different virtually autonomous universities. In Paris and other French cities no such umbrella structure exists, there we deal with completely autonomous universities which were originally part of one 'mother university'. As a consequence it is very cumbersome to distinguish between departments of these different universities within a city. The two 'Free Universities' of Brussels (Vrije Universiteit Brussel, VUB, and the Université Libre de Bruxelles, ULB) are a notorious example in this sense. Another well known problem is the 'interwovenness' of the French CNRS and French universities.

This problem is, in fact, a 'fine structure' problem: matching bibliometric data ('external') with the 'real fine structure' ('internal') of the principal organisation (e.g., a university). In order to do this, we need accurate 'fine-structure' data per organisation. Moreover, this internal structure is 'dynamic': new departments, schools, and certainly new research groups are created all the time.

I see at least two possibilities for tackling this problem. The first is the 'narrowing down of fields': *the smaller the bibliometric 'refinement' of fields* (e.g., from neuroscience as a whole to brain infarct research as a specific research theme within neuroscience), *the more we approach 'real' units* such as research groups within the internal structure of a principal organisation: 'convergence principle'. The bibliometric mapping methodology discussed in Section 3 (and a detailed discussion by Noyons (2004) in this handbook) is particularly suited to this approach.

A second approach concerns networks of co-operating scientists: the analysis of collaborating researchers provides the *internal structure* of that specific (sub-)field in terms of co-authors. Thus the real 'working floor' groups are identified (Vinkler, 1993; Melin and Persson, 1996; Glänzel, 2001). This identification is completely *independent* of the quality of information about principal organisation addresses. It is, as it were, based on a 'bibliometrically driven' *self-organisation* of science.

More generally, the understanding of research systems would benefit from the integration of bibliometric and other scientometric indicators into sociologically oriented studies (Gläser and Laudel, 2001).

2.3.4 Statistical issues: general ones and some related to journal impact

Standard statistical techniques relate to quantities that are distributed approximately 'normally'. Many characteristics of research performance, particularly those based on citation analysis, are not normally, but very skewly, distributed. Thus statistical averages can be misleading. For larger samples, such as the entire oeuvre of a research group over a period of years, the *central limit theorem* says that whatever the underlying distribution of a set of independent variables (provided that their variance is finite), the sum or average of a relatively large number of these variables will be a random variable with a distribution close to normal.

On the basis of these considerations I am confident that, for instance, our crown indicator ***CPP/FCSm*** does provide a useful measure. This can be proved empirically by the strong correlation of ***CPP/FCSm*** and the earlier discussed 'top 10%' indicator in which the distribution function is taken into account (Noyons et al., 2003).

A heavily debated theme in bibliometric studies is the 'predictive' character of journal impact, i.e., the relation between journal impact and the impact of a publication within that journal (see for instance Seglen, 1992, 1994; van Raan, 2001). In current research we focus in more detail on the relation between ***CPP*** and ***JCSm*** and other statistical characteristics of journal impact.

The indicators ***JCS*** and ***JCS/FCSm*** are novel journal indicators which characterise a journal in a more appropriate way than the commonly used

journal impact factors. The unique aspect of these journal impact indicators is that the type of paper (e.g., letters, normal article, review) *as well as* the specific years in which the papers were published are taken into account. This is absolutely necessary, as the average impact of journals may have considerable annual fluctuations and large differences per article type, see Moed and van Leeuwen (1995, 1996).

2.3.5 Peer review judgment and bibliometric findings....signs of theory–invariance?

The results of peer review judgment and those of bibliometric assessment are not completely independent variables. Peers take 'bibliometric aspects' into account in their judgment, for instance (number of) publications in the better journals. Thorough studies of larger-scale evaluation procedures in which empirical material is available with data on both peer judgment as well as bibliometric indicators are rare. I refer to Rinia et al. (1998) for a comparison of bibliometric assessment based on various indicators with peer review judgment in condensed matter physics, and to Rinia et al. (2001) for a study of the influence of interdisciplinarity on peer review in comparison with bibliometric assessment.

I have already mentioned the empirical gold mine we created with our long standing bibliometric practice. In current work the relation between bibliometric assessment and peer judgment for several hundreds of physics and chemistry research groups is studied. This is a unique collection of data. This study shows a striking agreement between elements of research performance measurement and the results of peer review. But at the same time remarkable differences are found in which not necessarily peer judgment has to be considered as 'right' (van Raan en van Leeuwen, 2004).

Indeed, peers may be right or wrong in their judgement. Also they undoubtedly use bibliometric elements in their judgement; for instance, they generally attach great value to publications in the top journals. Therefore, bibliometric findings and outcomes of peer review are not independent variables in the 'quality judgment space'. But this entanglement is unavoidable because (1) there is no higher authority to judge the quality of scientific work than a peer group of colleagues, and (2) attracting attention, provoking reactions by written communication, is very fundamental in most fields of science. *Any reasonable theory* has to 'accept this reality'. So if bibliometric analysisis is advanced in such a way that it becomes an indispensable instrument for measuring progress of science, and we think this stage is reached now, then we are approaching Holton's ideal of 'theory–invariant' indicators.

3. PRINCIPLES OF CONCEPT–SIMILARITY BASED MAPPING

Each year about a million scientific articles are published. How should one keep track of all these developments? Are there specific patterns 'hidden' in this mass of published knowledge at a 'meta level', and if so, how can these patterns be interpreted (Van Raan and Noyons, 2002)?

A first and crucial step is the definition of a research field. There are several approaches: on the basis of selected concepts (keywords) and/or classification codes in a specific database, selected sets of journals, a database of field–specific publications, or any combination of these approaches. Along these lines titles and abstracts of all relevant publications can be collected for a series of successive years, thus operating on many tens of thousands of publications per field. Next, with a specific computer-linguistic algorithm, titles and abstracts of all these publications can be parsed. This automated grammatical procedure yields all nouns and noun phrases (standardised) which are present in the entire set of collected publications (Noyons, 1999).

An additional algorithm creates a frequency list of these many thousands of parsed nouns and noun phrases while filtering out general, trivial words. The most frequent nouns/noun phrases can be considered as the most characteristic concepts of the field (this can be 100 to 1,000 concepts, say, N concepts). The next step is to encode each of the publications with these concepts. In fact this code is a binary string (yes/no) indicating which of the N concepts is present in title or abstract. This encoding is as it were the 'genetic code' of a publication. As in genetic algorithms, the encoding of each publication can be compared with that of any other publication by calculating pairwise the 'genetic code similarity' (here: concept similarity) of all publications in a specific field. The more concepts two publications have in common, the more these publications are related on the basis of concept similarity, and thus they can be regarded as belonging to the same sub-field, research theme, or research specialty. To use a biological metaphor: the more specific DNA elements two living beings have in common, the more they are related. Above a certain similarity threshold they will belong to a particular species.

The above procedure allows clustering of information carriers — the publications — on the basis of similarity in information elements — the concepts ('co-publication' analysis). Alternatively, the more specific concepts are mentioned together in different publications the more these concepts are related. Thus information elements are clustered ('co-concept' analysis). Both approaches, the co-publication and the co-concept analysis, are related by the rules of matrix algebra. In practice the co-concept

approach (Noyons and Van Raan, 1998) is most suited to science mapping, i.e., the 'organisation of science according to concepts'.

Intermezzo: For a supermarket 'client similarity' on the basis of shopping lists can be translated into a clustering either of the clients (information carriers, in which the information elements are the products on their shopping lists) or of the products. Both approaches are important: the first gives insight into groups of clients (young, old, male, female, different ethnic groups, etc.); and the second is important for the spatial division of the supermarket into product groups.

In outline the clustering procedure is as follows. First, for each field a matrix is constructed which composed of co-occurrences of the N concepts in the set of publications for a specific period of time. This 'raw co-occurrence' matrix is normalised in such a way that the similarity of concepts is no longer based on the pairwise co-occurrences but on the co-occurrence 'profiles' of the two concepts in relation to all other concepts. This similarity matrix is the input for a cluster analysis. Standard hierarchical cluster algorithm including statistical criteria can be used to find an optimal number of clusters. The identified clusters of concepts represent in most cases recognisable 'sub-fields' or research themes. Each sub-field represents a sub-set of publications on the basis of concept–similarity profiles. If any of the concepts is in a publication, this publication will be attached to the relevant sub-field. Thus publications may be attached to more than one sub-field. This overlap between sub-fields in terms of joint publications is used to calculate a further co-occurrence matrix, now based on sub-field publication similarity.

To construct a map of the field, the sub-fields (clusters) are positioned by multi-dimensional scaling. Thus sub-fields with a high similarity are positioned in each other's vicinity, and sub-fields with low similarity are distant from each other. The size of a sub-field (represented by the surface of a circle) indicates the share of publications in relation to the field as a whole. A two-dimensional structure is not sufficient to cover all relations embedded in the underlying matrix. Particularly strong relations between two individual sub-fields are indicated by a connecting line.

A next step (Noyons et al., 1999) is the integration of mapping *and* performance assessment. It enables us to position actors (such as universities, institutes, R&D divisions of companies, research groups) on the worldwide map of their field, and to measure their influence in relation to the impact-level of the different sub-fields and themes. Thus a strategic map is created: who is where in science, and how strongly?

A series of maps of successive time periods reveals trends and changes in structure, and even may allow 'prediction' of near-future developments by

extrapolation. Such changes in maps over time (field structure, position of actors) may indicate the impact of R&D programmes, particularly in research themes around social and economic problems. In this way our mapping methodology is also applicable in the study of the socio-economic impact of R&D.

Bibliometric maps provide an instrument which can be used optimally in an electronic environment. Moreover, there is a large amount of detailed information 'behind the maps'. Hence it is of crucial importance that this underlying information, particularly about research performance, can be retrieved in an efficient way, to provide the user with a possibility of exploring the fields and of judging the usefulness of maps against the user's own expertise. Advanced internet–based user-interface facilities are necessary (Noyons, 1999; Noyons, 2004, in this Handbook) to enable this further exploration of the maps and of the data 'behind the maps'. Thus bibliometric maps and their internet–based user-facilities will enable users to compare the scientific performance of groups/institutes with other 'benchmark' institutes. Likewise, the maps can be used for the selection of benchmark institutes, for instance institutes chosen by the experts.

Co-citation analysis provides an alternative type of mapping, but it unavoidably depends on the availability of citation (reference) data and thus its applicability is less general than concept–similarity mapping. Co-citation maps are based on the number of times two particular articles are cited together in other articles. The development of this analytical technique is based on the pioneering work of Henry Small (Small, 1973; Small and Sweeney, 1985; Small et al., 1985). When aggregated to larger sets of publications, co-citation maps indicate clusters of related scientific work (i.e., based on the same publications, as far as reflected by the cited literature). These clusters can often be identified as 'research specialties' (McCain, 1990; Bayer et al., 1990; White and McCain, 1998; Small, 1999; Prime et al., 2002). Their character may, however, be of a different kind compared with co-word based clusters: because they are based on citation practices they may reflect cognitive as well as social networks and relations (Braam et al., 1991a,b). Moreover, citations only reflect a part of the intellectual structure, and they are subject to a certain, often field–specific, time lag. For recent work on co-citation analysis for mapping research themes of socio-economic importance I refer to Schwechheimer and Winterhager (2001).

As Derek de Solla Price formulated twenty five years ago: "scientific papers themselves form a system with a visible structure and, indeed, one that appears highly deterministic: the universe of scientific papers exhibits a clustering structure in a space of surprisingly small dimensionality: most of the behaviour can be accounted for in the usual two dimensions of a

geographical map. The clusters correspond remarkably well to entities that we intuitively feel to be the basic sub-fields of which science is composed. Whatever their physical reality, maps of science are certainly useful as heuristic tools." (Price, 1978).

Mapping of science is a fascinating endeavour. For a detailed discussion of important new developments in bibliometric mapping I refer to the contribution of Noyons (2004) in this Handbook.

4. CONCLUDING REMARKS AND OUTLOOK

The quantitative study of science aims at the advancement of our knowledge on the development of science, also in relation to technological and socio-economic aspects. Bibliometric methods play an important role in this field of research. The field is both problem oriented as well as basic in nature. There are important interdisciplinary links with philosophy, history and sociology of science, with policy and management studies, with mathematics and physics, and particularly with information science.

I distinguish four inter-related research themes: (1) the development of methods and techniques for the design, construction, and application of quantitative indicators on important aspects of science; (2) the development of information systems about science; (3) the study of the interaction between science and technology; and (4) the study of cognitive and socio-organisational processes in the development of scientific fields.

The work in the *first* research theme concerns empirical studies on the assessment of research performance and directly related aspects such as publication and citation behaviour, notions of scientific quality, differences in communication practices in the different disciplines, comparison with qualitative judgments by peers. Standardisation of indicators including analysis of citing papers to assess aspects of 'knowledge users' mark the development of the 'second generation' bibliometric analysis (Van Leeuwen, 2004). At the same time it will be of crucial importance to monitor the influence of the various forms of electronic publishing on all bibliometric indicators, ranging from the mere number of publications to composed indicators such as the internationally normalised impact.

It is interesting to notice that only recently, owing to the gradually increasing number of applications of large-scale bibliometric analysis for research performance assessment, bibliometric characteristics of 'real' working floor entities such as research groups become known. So far, these characteristics have mainly concerned 'standard entities' such as authors, journals, universities, and countries. The study of the 'real working floor' enables the inclusion of further input data about personnel which goes

beyond the data which are strictly necessary for conducting the bibliometric analysis described in Section 2.2. For instance, data about the sex and age of researchers enables one to investigate the role of women (Lewison, 2001; Prpić, 2002) or of the different age categories in the science system.

We have emphasised in this contribution the potential of advanced bibliometric indicators as 'theory–invariant' measures of scientific progress. Nevertheless, in the application of bibliometric indicators, no matter how advanced, it will remain of the utmost importance to know the limitations of the method and to guard against misuse, exaggerated expectations of non-expert users, and undesired manipulations by scientists themselves (Adam, 2002; Butler, 2003; Weingart, 2003; Glänzel and Debackere, 2003).

Given the crucial role of data as building blocks for indicators, it is not a surprise that a considerable part of the research in the field is devoted to the *second* theme: the development and maintenance of science information systems. These systems may contain data of many millions of scientific publications, but equally important are the many methodological and technical 'added values'. This part of quantitative studies of science is mainly system design and software development, in order to handle the enormous data system and to apply complex algorithms for the calculation of a wide range of indicators, including new journal impact measures. In addition, other than the 'classic' bibliometric data may be added to enrich the system with, for instance, input data of scientific institutions and business companies, patent data, and web-based data (Björneborn and Ingwersen, 2001; Bar-Ilan, 2001; Thelwall and Harries, 2003). Here we have an interdisciplinary bridge to information and computer science.

In the *third* research theme the focus is on the interaction between science and technology. I mention as an example the study of author–inventor relations (i.e., scientists who are active both in writing research publications as well as in creating technological breakthroughs), and the use of scientific knowledge in technological innovations (Schmoch, 1993) on the basis of citation relations between patents and publications (Albert et al., 1991; Narin, 1994; Narin et al., 1997; Glänzel and Meyer, 2003). Technology in its turn strongly influences scientific progress (Etzkowitz and Leydesdorff, 2000), particularly by the ever advancing development of instruments and facilities. Therefore the study of the interaction between science and technology has to take a broader perspective than only the transfer of knowledge from science to the technological domain. Most probably the development of instruments is *the* driving force of science. Hence the development of indicators describing the 'instrumental state-of-the-art' in scientific fields is very important.

The *fourth* theme is strongly related to bibliometric mapping techniques. The central issue here is to find optimal visual representations at different

aggregation levels by exploring the idea of 'self-organising structures' in scientific and technological (on the basis of patents) development. It is a challenge to identify 'hidden patterns' in the enormous amount of data because all these publications (and patents) are connected by common references, concepts, classification codes. Co-citation and co-word techniques are examples of approaches to unravelling this gigantic network of inter-related pieces of scientific knowledge. These are important steps toward imaging cognitive processes. Systematic comparison of *cognitive* structures with *communication* structures based on citation analysis (Van Raan and Noyons, 2002) offers the possibility of discovering areas of science which are cognitively related but not connected in terms of reference practices (pioneering work by Swanson, 1986 and 1987).

Maps of science, with the locations of the major actors, are specific representations of scientific activities. They have practical values ('strategic overviews') as well as more cognitive (e.g., what type of scientific activities are primarily represented on the map). Co-word (concept similarity based) clusters can be used as *'journal set' independent* entities for defining (sub-) fields and research themes. An important advance in mapping is 'real time' user–driven application. This enables us to observe how differences in the definitions of fields (in terms of keywords, journals, etc.) lead to different maps, and, particularly, which defining elements really do matter. It also allows simulations and other manipulations that may teach us more about the meaning of science maps. This real-time mapping is absolutely necessary for making the next step: to know more about the relation between *cognitive* and *bibliometric* mapping.

Finally, an exciting development is the study of statistical and topological properties of bibliometric networks and their relation to other networks. Theoretical work is oriented towards the understanding of fractal properties of science as a 'bibliometric structure' in general, and of co-occurrence structures such as found in maps based on co-citation analysis in particular. Most probably these properties are related to (cumulative) growth phenomena (van Raan, 1990, 2000b). Soon the mapping and the network–based approaches will amalgamate. Bibliometric analysis then will reach its ultimate goal: to become, in the first place, an instrument for a scientist as a *grateful user*, instead of an instrument for a scientist as a *vulnerable target*.

To conclude this contribution, it is now not too vain to answer Holton's major question 'Can science be measured?' with a modest 'yes'.

REFERENCES

Adam, D. (2002). The counting house. *Nature*, 415, 726–729.

Albert, M.B., Avery, D., Narin, F., MacAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20, 251–259.

Arunachalam, S., Srinivasan, R., Raman, V. (1994). International collaboration in science-participation by the Asian giants. *Scientometrics*, 30, 7–22.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes — A review and analysis. *Scientometrics*, 50, 7–32.

Bayer, A.E., Smart, J.C., McLaughlin, G.W. (1990). Mapping intellectual structure of a scientific subfield through author cocitations. *Journal of the American Society for Information Science*, 41, 444–452.

Beaver, D. de B., Rosen, R. (1978). Studies in scientific collaboration, 1: Professional origins of scientific co-authorship. *Scientometrics*, 1, 65–84.

Björneborn, L., Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50, 65–82.

Borgman, C.L. (ed.) (1990). *Scholarly Communication and Bibliometrics*. Newbury Park: Sage.

Braam, R.R., Moed, H.F., van Raan, A.F.J. (1991a). Mapping of science by combined co-citation and word analysis, I: Structural Aspects. *Journal of the American Society for Information Science (JASIS)*, 42, 233–251, and II: Dynamical Aspects. *Journal of the American Society for Information Science (JASIS)*, 42, 252–266.

Braun, T., Glänzel, W., Schubert, A. (1988). World flash on basic research — The newest version of the facts and figures on publication output and relative citation impact of 100 countries 1981–1985. *Scientometrics*, 13, 181–188.

Braun, T., Glänzel, Grupp, H. (1995). The scientometric weight of 50 nations in 27 science areas, 1989–1993. 1: All fields combined, mathematics, engineering, chemistry and physics. *Scientometrics*, 33, 263–293; and 2: Life sciences. *Scientometrics*, 34, 207–237.

Brooks, T.A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science*, 37, 34–36.

Butler, L. (2003). Modifying publication practices in response to funding formulas. *Research Evaluation*, 17, 39–46.

Callon, M., Bauin, S., Courtial, J.P., Turner, W. (1983). From translation to problematic networks: an introduction to co-word analysis. *Social Science Information*, 22, 191–235.

Cole, S., Cole, J.R., Dietrich, L. (1978). Measuring the cognitive state of scientific disciplines. In: Elkana et al., op. cit.

de Candolle, A. (1873, 2nd. edition 1885). *Histoire des sciences et des savants depuis deux siècles*. Genève/Basel: H.Georg. Reprint in 1987 by Fayard.

Egghe L., Rousseau, R. (2000). The influence of publication delays on the observed aging distribution of scientific literature. *Journal of the American Society for Information Science*, 51, 158–165.

Elkana, Y., Lederberg, J., Merton, R.K., Thackray, A., Zuckerman, H. (Eds.) (1978). *Toward a metric of science: the advent of science indicators*. New York: John Wiley.

Etzkowitz, H., Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, 29, 109–123.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1, 359–375.

Garfield, E. (1980). Premature discovery or delayed recognition — Why? *Current Contents,* 21, May 26, 5–10.
Gilbert, G.N. (1978). Measuring the growth of science– review of indicators of scientific growth. *Scientometrics,* 1, 9–34.
Glänzel, W. (1996). A bibliometric approach to social sciences, national research performances in 6 selected social science areas, 1990–1992. *Scientometrics,* 35, 291–307.
Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics,* 51, 69–115.
Glänzel, W., Meyer, M. (2003). Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations. *Scientometrics,* 58, 415–428.
Glänzel, W., Schlemmer, B., Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics,* 58, 571–586
Glänzel, W., Debackere, K. (2003). *On the opportunities and limitations in using bibliometric indicators in a policy relevant context.* In: Bibliometric analysis in science and research. Applications, Benefits and Limitations. Second Conference of the Central Library, Forschungszentrum Jülich, (pp. 225– 236). (ISBN 3-89336-334-3).
Gläser, J., Laudel, G. (2001). Integrating scientometric indicators into sociological studies: methodical and methodological problems. *Scientometrics,* 52, 411–434.
Grupp, H., Schmoch, U., Hinze, S. (2001). International alignment and scientific regard as macro-indicators for international comparisons of publications. *Scientometrics,* 51, 359–380.
Haitun, S.D. (1982). Stationary scientometric distributions. 1: Different approximations. *Scientometrics,* 4, 89–104.
Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics,* 44, 193–215.
Holton, G. (1978). *Can science be measured?* In: Elkana et al., op. cit.
Horrobin, D.F. (1990). The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association (JAMA),* 263, 1438–1441.
Kamerlingh Onnes, H. (1882). *De betekenis van kwantitatief onderzoek in de natuurkunde* (The meaning of quantitative research in physics). Inaugural Address as Professor of Physics, Leiden University.
Koenig, M.E.D. (1983). Bibliometric indicators versus expert opinion in assessing research performance. *Journal of the American Society for Information Science,* 34, 136–145.
Kostoff, R.N. (1995). Federal research impact assessment — Axioms, approaches, applications. *Scientometrics,* 34, 163–206.
van Leeuwen, Th.N., Moed, H.F., Tijssen, R.J.W., Visser, M.S., van Raan, A.F.J. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics,* 51, 335–346.
van Leeuwen, Th.N. (2004). *Second generation bibliometric analysis.* Ph.D. Thesis Leiden University.
Lewison, G. (2001). The quantity and quality of female researchers: a bibliometric study of Iceland. *Scientometrics,* 52, 29–43.
Lewison, G. (2002). Researchers' and users' perceptions of the relative standing of biomedical papers in different journals. *Scientometrics,* 53, 229–240.
Lotka, A.J. (1926). The frequency distribution of scientific productivity. *J. Washington Acad. Sci.,* 16, 317–323.
MacRoberts, M.H., MacRoberts, B.R. (1996). Problems of citation analysis. *Scientometrics,* 36, 435–444.

MacRoberts, M.H., MacRoberts, B.R. (1988). Author motivation for not giving citing influences — A methodological note. *Journal of the American Society for Information Science,* 39, 432–433.
Martin, B.R., Irvine, J. (1983). Assessing basic research: some partial indicators of scientific progress in radio astronomy. *Research Policy,* 12, 61–90.
May, R.M. (1997). The scientific wealth of nations. *Science,* 275, 793–796.
McCain, K.W. (1984). Longitudinal author cocitation mapping — The changing structure of macroeconomics. *Journal of the American Society for Information Science,* 35, 351–359.
McCain, K.W. (1990). Mapping authors in intellectual space — A technical overview. *Journal of the American Society for Information Science,* 41, 433–443.
Melin, G., Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics,* 36, 363–377.
Moed, H.F., van Leeuwen, Th.N. (1995). Improving the accuracy of the Institute for Scientific Information's Journal Impact Factors. *J. of the American Society for Information Science (JASIS),* 46, 461–467.
Moed, H.F., van Leeuwen, Th.N. (1996). Impact factors can mislead. *Nature,* 381, 186.
Moed, H.F., Luwel, M., Nederhof, A.J. (2002). Towards research performance measurement in the humanities. *Library Trends,* 50, 498–520.
Moravcsik, M.J. (1975). Phenomenology and models of growth of science. *Research Policy,* 4, 80–86.
Moravcsik, M.J., Murugesan, P. (1979). Citation patterns in scientific revolutions. *Scientometrics,* 1, 161–169.
Moxham, H., Anderson, J. (1992). Peer review. A view from the inside. *Science and Technology Policy,* February 1992, 7–15.
Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity.* Washington D.C.: National Science Foundation.
Narin, F. (1978). Objectivity versus relevance in studies of scientific advance. *Scientometrics,* 1, 35–41.
Narin, F. (1994). Patent bibliometrics. *Scientometrics,* 30, 147–155.
Narin, F., Hamilton, K.S., Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy,* 26, 317–330.
National Science Board (1973). *Science Indicators 1972.* Washington DC: Government Printing Office.
Nederhof, A.J. (1988). *The validity and reliability of evaluation of scholarly performance.* In: A.F.J. van Raan (ed). (1988), Handbook of Quantitative Studies of Science and Technology (pp.193–228). Amsterdam: Elsevier/North-Holland, (ISBN 0-444-70537-6).
Noma, E. (1982). An improved method for analysing square scientometric transaction matrices. *Scientometrics,* 4, 297–316.
Noyons, E.C.M., van Raan, A.F.J. (1998). Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research. *J. of the American Society for Information Science and Technology (JASIST),* 49, 68–81.
Noyons, E.C.M., Luwel, M., Moed, H.F. (1999). Combining mapping and citation analysis for evaluative bibliometric purpose. A bibliometric study on recent development in micro–electronics. *Journal of the American Society for Information Science and Technology (JASIST),* 50, 115-131.
Noyons, E.C.M. (1999). *Bibliometric mapping as a science policy and research management tool.* Ph.D. Thesis Leiden University. Leiden: DSWO Press (ISBN 90-6695-152-4).
Noyons, E.C.M., Buter, R.K., van Raan, A.F.J., Schmoch, U., Heinze, T., Hinze, S., Rangnow, R. (2003). *Mapping excellence in science and technology across Europe* (Part

1: Life sciences, Part 2: Nanoscience and nanotechnology). Report to the European Commission by the Centre for Science and Technology Studies (CWTS), Leiden University, and the Fraunhofer Institute for Systems and Innovation Research (Fraunhofer-ISI), Karlsruhe.

Noyons, E.C.M. (2004). *Science Maps within in a Science Policy Context.* This Handbook.

OECD (1963). *The measurement of scientific and technological activities,* 'Frascati Manual', Paris: Organization for Economic Co-operation and Development (OECD).

Peritz, B.C. (1983). A classification of citation roles for the social sciences and related fields. *Scientometrics,* 5, 303–312.

Porter, A.L., Chubin. D.E. (1985). An indicator of cross-disciplinary research. *Scientometrics,* 8, 161–176.

De Solla Price, D.J. (1978). Toward a model for Science Indicators. In: Elkana et al., op. cit.

De Solla Price, D.J. (1981). The analysis of scientometric matrices for policy implications. *Scientometrics,* 3, 47–53.

Prime, C., Bassecoulard, E., Zitt, M. (2002). Co-citations and co-sitations: A cautionary view on an analogy. *Scientometrics,* 54, 291–308.

Prpić, K. (2002). Gender and productivity differentials in science. *Scientometrics,* 55, 27–58.

van Raan, A.F.J. (ed). (1988). *Handbook of Quantitative Studies of Science and Technology.* Amsterdam: Elsevier/North-Holland (ISBN 0-444-70537-6).

van Raan, A.F.J. (1990). Fractal dimension of co-citations. *Nature,* 347, 626.

van Raan, A.F.J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics,* 36, 397–420.

van Raan, A.F.J. (1997). Scientometrics: State-of-the-Art. *Scientometrics,* 38, 205–218.

van Raan, A.F.J. (1998). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics,* 43, 129–139.

van Raan, A.F.J. (2000a). The Pandora's box of citation analysis: measuring scientific excellence, the last evil? In: B. Cronin and H. Barsky Atkins (eds.). *The Web of Knowledge. A Festschrift in honor of Eugene Garfield.* Ch. 15, p. 301–319. Medford (New Jersey): ASIS Monograph Series, 2000 (ISBN 1-57387-099-4).

van Raan, A.F.J. (2000b). On growth, ageing, and fractal differentiation of science. Scientometrics 47, 347–362.

van Raan, A.F.J. (2001). Two–step competition process leads to quasi power–law income distributions. Application to scientific publication and citation distributions. *Physica A,* 298, 530–536.

van Raan, A.F.J., Noyons, E.C.M. (2002). *Discovery of patterns of scientific and technological development and knowledge transfer.* In W. Adamczak, A. Nase (Eds.), Gaining Insight from Research Information. Proceedings of the 6th International Conference on Current Research Information Systems, University of Kassel, August 29–31, 2002 (pp. 105– 112). Kassel: University Press, (ISBN 3-933146-844).

van Raan, A.F.J., van Leeuwen, Th.N. (2002). Assessment of the scientific basis of interdisciplinary, applied research. Application of bibliometric methods in nutrition and food research. *Research Policy,* 31, 611–632

van Raan, A.F.J. (2003). *Reference–based publication networks with episodic memories.* E–print ArXiv cond-mat/0311318.

van Raan, A.F.J. (2004). Sleeping Beauties in Science. *Scientometrics,* 59, 461–466.

van Raan, A.F.J., van Leeuwen, Th.N. (2004). *Statistical aspects of research group performance, journal impact, and peer judgement.* To be published.

Rinia, E.J., van Leeuwen, Th.N., van Vuren, H.G., van Raan, A.F.J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria. Evaluation of condensed matter physics in the Netherlands. *Research Policy,* 27, 95–107.

Rinia, E.J., van Leeuwen, Th.N., van Vuren, H.G., van Raan, A.F.J. (2001). Influence of interdisciplinarity on peer–review and bibliometric evaluations. *Research Policy,* 30, 357–361.

Rip, A., Courtial, J.P. (1984). Co-word maps of biotechnology — An example of cognitive scientometrics. *Scientometrics,* 6, 381–400.

Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics,* 26, 193–211.

Schwechheimer, H., Winterhager, M. (2001). Mapping interdisciplinary research fronts in neuroscience: a bibliometric view to retrograde amnesia. *Scientometrics,* 51, 311–318.

Schubert A., Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientometric publications. *Scientometrics,* 5, 59–74.

Seglen, P.O. (1992). The skewness of science. *Journal of the American Society for Information Science,* 43, 628–638.

Seglen, P.O. (1994). Causal relationship between article citedness and journal impact. *Journal of the American Society for Information Science,* 45, 1–11.

Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Publications. *Journal of the American Society for Information Science,* 24, 265–269.

Small, H., Greenlee, E. (1980). Citation context analysis of a co-citation cluster- recombinant DNA. *Scientometrics,* 2, 1980.

Small, H., Sweeney, E. (1985). Clustering the Science Citation Index using co-citations, I: A Comparison of Methods. *Scientometrics,* 7, 393–404.

Small, H., Sweeney, E., Greenlee, E. (1985). Clustering the Science Citation Index using co-citations, II: Mapping Science. *Scientometrics,* 8, 321–340.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science,* 50, 799–813.

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine,* 30, 7–18.

Swanson, D.R. (1987). Two medical literatures that are logically but not bibliographically connected. *Journal of the American Society for Information Science,* 38, 228–233.

Sullivan D., Koester, D., White, D.H., Kern, R. (1980). Understanding rapid theoretical change in particle physics- a month-by-month co-citation analysis. *Scientometrics,* 2, 309–319.

Thelwall, M., Smith, A. (2002). Interlinking between Asia-Pacific University Web sites. *Scientometrics,* 55, 363–376.

Thelwall, M., Harries, G. (2003). The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society for Information Science,* 54, 594–602.

Vinkler, P. (1993). Research contribution, authorship and team cooperativeness. Scientometrics 26, 213–230.

Vinkler, P. (1998). Comparative investigation of frequency and strength of motives toward referencing, the reference threshold model- comments on theories of citation? *Scientometrics,* 43, 107–127.

Vláchy, J. (1979). Mobility in science. Bibliography of scientific career migration, field mobility, international academic circulation and brain drain. *Scientometrics,* 1, 201–228.

Weingart, P. (2003). *Evaluation of research performance: the danger if numbers.* In: Bibliometric analysis in science and research. Applications, Benefits and Limitations. Second Conference of the Central Library, Forschungszentrum Jülich (pp. 7–19). (ISBN 3-89336-334-3).
White, H.D., Griffith, B.C. (1981). Author cocitation- a literature measure of intellectual structure. *Journal of the American Society for Information Science,* 32, 163–171.
White, H.D., McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science,* 49, 327–355.
Wouters, P.F. (1999), The Citation Culture, PhD thesis, University of Amsterdam.
Ziman, J. (1978). *From Parameters to Portents -and Back.* In: Elkana et al., op.cit.

Chapter 2

ECONOMETRIC APPROACHES TO THE ANALYSIS OF PRODUCTIVITY OF R&D SYSTEMS

Production Functions and Production Frontiers

Andrea Bonaccorsi[1] and Cinzia Daraio[2]

[1] *School of Engineering, University of Pisa, Italy*
[2] *Institute for Informatics and Telematics – CNR and Sant'Anna School of Advanced Studies, Pisa, Italy. E-mail: cinzia@sssup.it*

Abstract: In this chapter we review and discuss the potential and limitations of econometric methods for the evaluation of productivity of scientific and technological (S&T) systems. We examine and compare the main approaches that have been applied in the literature: the production function and the production frontier approach. Both approaches present advantages and disadvantages. In the first part of the chapter we carry out a selective review of the two fields. In the second part we focus on the last developments of the efficiency analysis literature, with particular attention to the nonparametric approach. An illustration of the potential of robust nonparametric techniques is offered using data from the Italian National Research Council (CNR). The chapter concludes by discussing the potential of these approaches for the analysis of S&T systems beyond the existing applications.

1. INTRODUCTION

In this chapter we review and discuss the potential and limitations of econometric methods for the evaluation of productivity of S&T systems.

Any notion of productivity relates a vector of inputs to a vector of outputs. Unfortunately, in S&T systems all three definitional elements of productivity (inputs, outputs and the functional relation between the two) are affected by severe conceptual and measurement problems.

S&T production is based on a multi-input, multi-output relation, in which, differently from standard production activity, both inputs and outputs

Rousseau, S., Rousseau, R. (1997). Data envelopment analysis as a tool for constructing scientometric indicators. *Scientometrics*, 40, 45–56.
Rousseau, S., Rousseau, R. (1998). The scientific wealth of European nations: taking effectiveness into account. *Scientometrics*, 42, 75–87.
Sarrico, C.S., Hogan, S.M., Dyson, R.G., Athanassopoulos, A.D. (1997). Data envelopment analysis and university selection. *Journal of the Operational Research Society*, 48, 1163–1177.
Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
Simar, L. (2003). How to improve the performance of DEA/FDH estimators in the presence of noise? Discussion Paper to appear, Institut de Statistique, UCL, Belgium.
Simar, L., Wilson, P.W. (1999). Estimating and bootstrapping Malmquist indices. *European Journal of Operational Research*, 115, 459–471.
Simar, L., Wilson, P.W. (2000). Statistical inference in nonparametric frontier models: The State of the Art. *The Journal of Productivity Analysis*, 13, 49–78.
Simar, L., Wilson, P.W. (2001). Testing restrictions in nonparametric efficiency models. *Communications in Statistics*, 30 (1), 159–184.
Simar, L., Wilson, P.W. (2002). Nonparametric tests of returns to scale. *European Journal of Operational Research*, 139, 115–132.
Simar, L., Wilson, P. (2003a). Estimation and inference in two-stage, semiparametric models of production processes. Discussion Paper no. 0307, Institut de Statistique, UCL, Belgium.
Simar, L., Wilson, P. (2003b). Efficiency analysis: the statistical approach, Lectures Notes. Institute of Statistics, UCL, Belgium.
Simar, L., Zelenyuk, V. (2003). Statistical inference for aggregates of Farrell–type efficiencies. Discussion Paper no. 0324. Institut de Statistique, UCL, Belgium.
Stephan, P.E. (1996). The economics of science. *Journal of Economic Literature*, 34, 1199-1235.
Stephan, P.E., Levin, G. (1996). The critical importance of careers in collaborative scientific research, *Revue d'économie Industrielle*, 79 (1), 45-61.
Thanassoulis, E., Dunstan, P. (1994). Guiding schools to improved performance using data envelopment analysis: an illustration with data from a local education authority. *Journal of the Operational Research Society*, 45 (11), 1247–1262.
Thursby, J. (2000). What do we say about ourselves and what does it mean? Yet another look at economics department research. *Journal of Economic Literature*, 38 (2), 383–404.
Thursby, J., Kemp, S. (2002). Growth and productive efficiency of university intellectual property licensing. *Research Policy*, 31, 109–124.
Thursby, J., Thursby, M. (2002). Who is selling the ivory tower? Sources of growth in university licensing. *Management Science*, 48(1), 90–104.
Verry D.W., Layard, P.R. (1975). Cost functions for university teaching and research, *Economic Journal*, 85, 55–74.
Verry, D.W., Davis, B. (1976). University costs and outputs. Amsterdam: Elsevier.
Zhang, Y., Bartels, R. (1998). The effect of sample size on mean efficiency in DEA with application to electricity distribution in Australia, Sweden and New Zealand. *The Journal of Productivity Analysis*, 9, 187–204.
Zhu, J. (1996). Data envelopment analysis with preference structure. *Journal of the Operational Research Society*, 47, 136–150.

Chapter 3

INDICATORS FOR NATIONAL SCIENCE AND TECHNOLOGY POLICY

Their Development, Use, and Possible Misuse

Hariolf Grupp[1] and Mary Ellen Mogee[2]

[1]*Karlsruhe University and Fraunhofer ISI, Karlsruhe, Germany*
E-mail: h.grupp@isi.fraunhofer.de
[2]*Mogee Research & Analysis, Reston, Virginia, USA*

Abstract: The purpose of this chapter is to present a survey of the development of Science and Technology (S&T) indicators and their use in national policy making as well as to provide evidence of the vulnerability of S&T indicators to manipulation. A brief history of the development of S&T indicators begins with the United States followed by their worldwide diffusion, with particular emphasis on Europe. The current status of S&T indicators and newer developments towards composite indicators, benchmarking, and scoreboarding is discussed. To investigate the robustness of innovation scoreboards empirically a sensitivity analysis of one selected case is presented. It is shown that composite scores and rank positions can vary considerably, depending on the selection process. It seems not to be too difficult to argue for a 'country friendly' selection and corresponding weighting of indicators. Thus the use of scoreboards opens space for manipulation in the policymaking system. Further research is needed on alternative methods of calculation to prevent their misuse and abuse.

1. INTRODUCTION

The purpose of this chapter is to present a survey of the development of science and technology (S&T) indicators and their use in national policy making, to provide evidence of the vulnerability of S&T indicators to

manipulation and to suggest questions for future research.[1] The chapter is organised as follows. In the second section we review briefly the history and development of S&T indicators and the major S&T indicators reports in the United States, Japan, and Europe. The third section presents a sensitivity analysis of a composite indicator. Section 4 presents our conclusions and suggestions for future research.

2. A BRIEF HISTORY OF THE DEVELOPMENT OF S&T INDICATORS

2.1 The Beginning — the United States

Science and technology indicators had their start in the United States. The first *Science Indicators* report was published in 1973. The National Science Board (NSB), the policymaking board of the National Science Foundation (NSF), was mandated by the US Congress to publish the report biennially. (Perhaps not surprisingly, given the NSF's pre-eminent support for basic scientific research, it was not until 1987 that the focus and title of the report were broadened to *Science and Engineering Indicators*). In practice the report has been prepared by the Science and Engineering Indicators Unit (SIU) in the NSF's Science Resources Studies division and reviewed by the NSB, which also prepares a brief discussion piece as part of the report.

The idea of 'science indicators' (SI) was an outgrowth of the move toward 'social indicators', i.e., indicators similar to economic indicators such as Gross National Product (GNP) that would provide measures of the health of society. Based on the model of economic indicators, some observers expected science indicators to be a narrow range of statistics that tells about a larger universe, in the same way that new housing starts tell how the economy in general is doing. To date the science indicators report has not evolved in this direction, but rather has become a compendium of many different statistics to measure the health of US science and technology and to compare the US with other nations.

In the early years there was considerable criticism of the science indicators report including the publication– and patent–based indicators. The NSB member and famous mathematician Saunders MacLane criticised the

[1] An earlier, shorter version of this chapter was presented at the Conference in honour of Keith Pavitt, University of Sussex, November 13th–15th, 2003, and subsequently published in a partly overlapping way in a Special Issue of the journal 'Research Policy'.

publication–based indicators, in particular. Early SI reports counted publications by country instead of by individual's addresses or institution, which resulted in the finding that one third were British. In response to the critical writing the NSF held a world conference on the coverage and validity of the set of journals and the way of counting publications.

Similarly there was considerable scepticism in the beginning about patents as science or technology indicators. The NSF had to do lots of studies to show that the use of patents had methodological backing. However, it was patent indicators which showed the US that it should pay more attention to Japan as a competitive industrial power. In the part devoted to international comparisons increases were found in Japanese R&D funding and in all categories of patenting. This finding — that Japan was a power to contend with — was surprising to many at the time.

In addition to criticism of data and methodologies used in the SI report, some criticism of science indicators was based on resistance in the science community to making government funding decisions based on quantitative indicators. Many scientists believed, and continue to believe, that such decisions should be made on the basis of peer review.

In 1985 the House of Representatives Committee on Science and Technology undertook a Science Policy Study which asked the Office of Technology Assessment to examine "... the extent to which decision making would be improved through the use of quantitative mechanisms associated with the concept of investment". OTA concluded that "... while there are some quantitative techniques that may be of use to Congress in evaluating specific areas of research, basic science is not amenable to the type of economic analysis that might be used for applied research or product development". In his accompanying letter, John H. Gibbons, the OTA Director, stated further:

> "Much of the vitality of the American research system lies in its complex and pluralistic nature. Scientists, citizens, administrators, and Members of Congress all play various roles leading to final decisions on funding. While there may be ways to improve the overall process, reliance on economic quantitative methods is not promising. Expert analysis, openness, experience, and considered judgement are better tools."

Despite resistance, over time wide acceptance of the indicators grew. This has been owed in part to pressure from users and decision makers who wanted to be able to show that investment in S&T had value and impact. In response, the SI unit moved beyond the original indicators to additional indicators of interest to policy makers. There was constant improvement by taking feedback and criticism into account.

It was recognised early on that some indicators represent science and engineering resources or 'inputs' to the process of science and engineering, whilst others represented the process itself ('throughput' or 'flow'), the results or 'outputs' of science and engineering, and the effects or 'impacts' of science engineering. The early indicators tended to be heavy on inputs and throughput and weaker on outputs and impacts, spurring the SIU to look for and develop more indicators of outputs and impacts.

In addition to indicators of the level of scientific and technological activity, indicators were developed which examined international collaboration and intersectoral — e.g., university–industry — collaboration. Funding, citations, and co-authorship were all used as indicators of collaboration.

The SIU also developed indicators of the importance of science and engineering. For instance they looked at publication and patent. They also looked at patents citing literature (see, e.g. Narin and Noma, 1985). Publication citations have been used to indicate the quality of research. Patent citations have been used variously to indicate 'technological significance', 'social economic value' and 'private economic value' (for the many applications of patent analysis, see Breitzmann and Mogee, 2002). Other measures of value or quality from patent indicators include the rate at which patent maintenance fees are paid and the number of patent countries in which patent protection is sought for an invention (patent family size).

Later under the Government Performance and Results Act (GPRA) and the pressure to show use of government funded basic research, one of the key indicators that was helpful to policy makers was the patent citations to basic research. This was used to show that commercial technology was building on the basic research funded by government. Other agencies doing this today are the Office of Naval Research (ONR), the Department of Defense (DOD), the Department of Energy (DoE) and the Department of Veterans Affairs (DVA). Industry also uses patent citations and their references to basic research. It is difficult for funding agencies to show a direct use for the basic research they support. The Office of Management and Budget (OMB) and the Congress both review GPRA reports, and the OMB, in particular, likes numbers, so research funding agencies have used these numbers on patents citing basic research to argue for higher budgets.

The SI unit conducted a series of 4–5 surveys of the usage of the SI report. These were generally biennial, like the Science & Engineering report, and were sent out with the report. At first the questions were not split up by what specific indicators the people were using. Instead they asked about the policy issues the respondents wanted to cover and what they found useful. The users always asked for output indicators. Publications and patents are both output indicators. Therefore they had high priority for SI and they still

do. Responses were received from the NSB, scientific community, the Office of Science and Technology Policy (OSTP), OMB, and the OECD, amongst others. They all wanted the indicators to be relevant and useful.

The National Science Board (NSB) and the Director of the National Science Foundation (NSF) are important users of the publication data in the SI report. This is natural because publication indicators pertain largely to academic research, and the NSB and the NSF Director are major players in US science policy. This kind of information is reportedly used in allocating funds between scientific fields. Similarly, when they first found evidence of the high citation rate from patents to basic research, the NSB used it to make the industrial use of basic research the centerpiece of their discussion piece in the SI report.

That the NSB has continued the publication of the SI report, although it make suggestions for changes, suggests that the members find the compilation to be useful. In response to a recent (2003) proposal to cut back the SI report and to include just those statistics that are not published elsewhere, the NSB member Anita Jones stated in a presentation that:

- *Science & Engineering Indicators* (S&EI) is a leading data source for R&D policymakers;
- The data are sound; the definitions and categories change slowly;
- The longitudinal data emphasis is very useful;
- The data are collected in one place and repeatedly updated.

Dr. Jones went on to say that the audience for S&EI includes federal and state political appointees in the R&D area, scientists and engineers on advisory boards such as the NSB, the NSF directorate advisory committee, many levels of agency advisory committees, and National Academies' task forces – all important players in the US science policy system. In Dr. Jones' words, "S&EI is the 'one stop shop' for policy makers who do not study the multiplicity of R&D statistics publications." (July 2003)

The authors are not aware of any instrumental use of publication- or patent–based indicators in national policy, that is, cases in which a decision hinged upon the publication or patent indicators. There is wide recognition, however, that statistics and indicators are often used to justify decisions or to support a particular side in a disputed issue. It is also recognised that there is not necessarily anything wrong with using statistics in this manner, but the limitations and meaning of the data need to be made clear when they *are* used in this way. These types of use of S&T indicators present substantial potential for abuse.

Today the indicators in the S&EI report are in widespread use in the US. A broad range of participants in the policy process cites the statistics. Particularly since the federal government came under pressure from the

GPRA, S&T indicators have been reinvigorated in their use. They tend to be used to show where the US stands with respect to other countries and as a reason for increasing funding to particular areas of science or technology.

Also in the US, the Council on Competitiveness has developed its own set of innovation indicators. It uses patent data together with industry analysis. They are thinking about repeating their innovation summit. It is a lot of work and financially difficult in both publications and patents (particularly patents). It needs to be done in a collaborative way.

2.2 Worldwide Diffusion of S&T Indicators and the OECD

Today many other countries use the US data or have been inspired by the US to develop their own indicators and indicator systems. The US is still encouraging the spread of science indicators by consulting with countries which are establishing systems to track and use science indicators. The indicators are broadly accepted. For example, Latin America has become involved in S&T indicators. SI helped RICYT (La Red Iberoamericana de Indicadores de Ciencia y Tecnología) to start a S&T indicators network. Over the past 5–6 years they have developed comparable data such as the SEI pocketbook data book. It covers all the Americas, including the US and Canada, plus Spain and Portugal. It includes a broad range of indicators such as R&D funding, publications, and patents.

It may be an overstatement to say that other countries have gone further than the US in the use of indicators. However, it seems that European countries may be going in different directions in the use of S&T indicators, using indicators more in benchmarking and in foresight exercises than the US (see section 2.3 below).

The idea of developing a S&T indicator system that includes publications and patents and its use by policy makers is taken further in certain countries and used more rigidly in terms of funding individual researchers than it is in the US. Countries which have a more centralised science policy system tend to use science and technology indicators in a more rigorous way. For example, in France and Mexico quantitative indicators are used in the decision to give individual researchers more funding. In the US S&T indicators are used more widely by management for general policy and awareness of trends than at the level of allocating resources to individual researchers.

Beginning in the late 1970s the Organisation for Economic Co-operation and Development's (OECD) secretariat for science and technology indicators (restructured and renamed several times in the past decades) exerted a very important standardising role within the member states of the

OECD. By inviting researchers, statisticians, and other responsible persons to join workshops, by editing manuals on R&D, patent and innovation measurement, and revising unclear national statistics to OECD standards, the secretariat carried out an important task by making national scoreboards on S&T comparable. The OECD bodies resisted producing only simple aggregated rank tables of countries' innovation performance. Even the most recent STI scoreboard (OECD, 2003), the sixth in a biennial series which started a decade ago, did not produce scalar measures of innovation activities of countries, although it did give particular attention to offering new or improved official measures for international comparisons in the major areas of policy interests. This stands in clear distinction to what is observed at the European level, which will be reported on below.

2.3 S&T Indicators in Japan

In Japan, for reasons of language, major English reports on S&T and the respective indicators began with the establishment of NISTEP (National Institute of Science and Technology Policy) in the year 1988. Certainly, important Japanese sources on R&D expenditures were published before, such as the Report on the Survey of Research and Development (Management and Coordination Agency, various years) issued annually in Japanese with English sub-titles to table and figure captions. Yet the special dedication of the foundation of NISTEP was to bring Japanese S&T information, including indicators, to an international audience. Within the mission of NISTEP to contribute to policy making by taking a sort of task force, major internationally comparable reports on S&T indicators were published. These indicators systematically organised the knowledge about scientific and technological activities of Japan and the corresponding reports were the first to make the overall state of these activities quantitatively comprehensible.

In the context of this chapter, in focussing on aggregation methods of S&T indicators, the approach of NISTEP to establish a Japanese science and technology system must be recognised. The basic method of 'integrating' S&T indicators by the Japanese institution was in using a 'cascade model' and factor analysis (Niva and Tomizawa, 1995; Kodama, 1987). The international comparison of overall strengths in science and technology was processed in such a way that 13 indicators for Japan and other countries were used to illustrate national S&T activities such as inputs in R&D, staff, output, number of scientific paper citations, and so on. That is, the multiplicity of indicators was reduced in a way that looks for similarities in the structure of the data and results in a lower-dimensional array of

indicators, which is more than a simple ranking. To the best of our knowledge these activities were not fully continued in the past years.

2.4 S&T Indicators in Europe

In Europe reporting on national science and technology performance has changed markedly in the past ten years. There are two main reasons for this. First, the former communist countries did not keep with OECD conventions, and the little information which was available before around 1990 was often not comparable. It was also widely considered to be systematically overestimated. This started to change around 1990 and since then several national reports from Eastern European countries have been issued, some of them in the English language. Yet in this short chapter it is impossible to give a full account in this respect (see, for instance, Gokhberg et al., 1999; CSRS, 1998).

The second change was the more active European Commission. A landmark in this respect is the first European report on S&T indicators (European Commission, 1994), which consisted of a massive attempt to collect available data of various kinds. The Commission was assisted by a large group of leading European researchers in that area. In 2003 the third such report was issued. This new role of the European Commission triggered numerous competing activities which led some observers to note an 'oversupply' of S&T indicator reports (see below).

Before these two trends made themselves felt in the 1990s and in the first years of the 21st century, a variety of non-comparable reporting systems in major Western European countries was in place. Here again we do not attempt to give a full account of the 1970s and the 1980s. Nor is it possible to do justice to every country. Some of these reports were not published regularly but only in exceptional cases and in different formats and many of them are in national languages. These activities were not stopped when the European Union level came up with own products and most of the national series continue today.

To give a few examples let us mention the French report on 'Science & Technologie – Indicateurs' which has been published since the inception of the Observatoire des Sciences et de Technologie (OST) in Paris in 1990 (OST, various years). The series of reports is clearly subdivided into the national level, the European level, and the international level. Observers within France are proud of the many data series and the systematic and continuous way the report is published. However, the reports are known to lack analytic and policy sections and assessments to complement the data series.

In the United Kingdom various related publications exist, but periodical reports in a consistent format that provide comparable information over a longer interval were not established. Amongst the newer publications let us just mention the Department of Trade and Industry (DTI) economics paper No. 7, 'Competing in the global economy – the innovation challenge' (2003).

In Germany, since 1965 the 'Bundesbericht Forschung' (Federal Research Report) is published every four years in the German language (BMBF, various years). From time to time more or less abridged English versions are available. The latest such report appeared very recently (May 2004). Because of the federal structure in Germany, this report has a national part, a state (Länder) part, and also some international comparisons. It is more focussed on R&D inputs and R&D infrastructure, and describes large organisations in Germany. This report is assisted by the 'Bericht zur technologischen Leistungsfähigkeit' (report on technological competitiveness), which has been published annually since 1985 (with various editors; for the latest version see Grupp et al., 2003). The latter report is not as complete as the former in terms of compiling R&D data and it has a less official character. It is prepared by research institutes for the German government and is quite analytic and policy oriented. Government officials occasionally were not happy with the assessments and findings. In the case of Germany one can also demonstrate the problems with former communist countries. The very complete R&D statistics of East Germany (former German Democratic Republic) had to be adjusted in a very complicated way in order to be comparable to the Western system. This was done in the first years of unification and now comparable backward information is available. This case may be taken as typical for all the Eastern European countries.

National reports are available from Austria (Pohn-Weidinger et al., 2001; Republik Österreich, 2003), Italy, the Netherlands, the Scandinavian countries, and so forth. Most of them are in the national languages.

Returning to the European Union level, in the past several years, in addition to the three European reports of S&T indicators mentioned previously, a variety of other reporting systems were established. Benchmarking activities were started with the explicit aim of going beyond existing statistics and providing new types of data not available so far (for instance, R&D staff by gender; European Commission, 2002). These activities are co-ordinated by the Directorate General for Research and assisted by a High Level Group of Experts on Benchmarking, Excellence, Co-Ordination of National Policies. The Directorate General for Research also issued the booklet 'Key Figures 2003–2004' (2003b). A preliminary version of an 'Innovation Scoreboard' was published in 2000 by the

innovation/Small and Medium Sized Industry programme of the Enterprise Directorate General. Since then the European Innovation Scoreboard has been published regularly. The same directorate also published a Biotechnology Innovation Scoreboard (2003). Both types of reports make use of 'composite indicators', which we discuss in more detail below. Composite indicators are also part of chapter 1 in the most recent European Report on Science and Technology indicators (2003).

It seems that the European Commission is driving S&T indicators in the direction of aggregation of different types of indicators into simpler constructs in order to summarise complex multi-dimensional phenomena.

2.5 Current Status of S&T Indicators

To summarise their development, S&T indicators have evolved over the past 30 years to become a large number of statistics each of which describes a portion of the science and technology system. Because they are by definition partial they must be used in combination with each other and with other kinds of data such as expert opinion to provide a full picture. Every indicator has its own unique strengths and weaknesses, and indicators need to be selected according to the problem or question being addressed (Grupp, 1998).

Progress has been made toward linking particular indicators to particular parts of the S&T system or the innovation process. Researchers have moved from the simple concepts of inputs and outputs to concepts of inputs, throughputs, outputs, and impacts at *various stages of the process*. Publications can be used to indicate the output of basic scientific research, for example, but would be misleading if used (alone) to indicate the output of industrial research and technology development. Patents, on the other hand, are useful as an indicator of applied research and technology development (*loc. cit.*).

It is increasingly recognised that some indicators are appropriate in certain contexts and not in others. For example, if the objective is to understand the development of computer software in the last quarter of the twentieth century, a patent analysis would not be recommended. This is because, although software is increasingly patented, particularly in its early years it was not, so an analysis of software patents would miss much software activity in the early years. However, if one is interested in the extent to which ownership of software patents is concentrated in a few companies, an analysis of patents would make sense.

Progress has also been made in developing indicators of quality, importance, or value, although these concepts themselves have not been defined as well as they should be.

Given the above situation, S&T indicators can be misused or abused, as well as used for positive purposes (Pavitt, 1988). Some of the possible misuses include:

- Reliance on a single indicator;
- Use of an indicator that is inappropriate for the technology, system, or stage of the R&D process;
- Drawing conclusions which are too strong, given the 'indicative' nature of indicators;
- Making inferences that are inappropriate, based on the indicator and its relationship to the phenomenon of interest.

2.6 The Development of Composite Indicators and Related Concepts

To sum up several decades of debate, the measurement of science and technology requires measurements along many dimensions. To date no ideal 'catch all' variable for science or innovation has been developed (Patel & Pavitt, 1995). Therefore in many cases multiple indicators have been used. However, the use of multiple indicators means that conventional methods such as the knowledge production function (Griliches, 1995) and many other concepts of efficiency measurement cannot be supported. Optimal configurations of measurement must be worked out in some other way (such as factor or data envelopment analysis). The recognition of the need to measure multiple dimensions of science and technology has also led to the emerging and pioneering field of composite indicators to enlighten national S&T policies.

A fortiori, the multi-dimensional science and technology (S&T) variables are usually not expressed in monetary terms but rather are measured in other units (such as patent counts, innovation counts, number of citations, etc.) and may not be comparable to each other. Lacking a well defined correspondence between relevant S&T data — for instance a conversion relationship between dollars and patent numbers — the multi-dimensional profiles cannot be aggregated into an overall scalar figure. This situation is fundamentally different from one in which all variables are fully specified in terms of quantities and costs or prices such as such well known economic indicators as the Gross Domestic Product (GDP).

On the micro-level of companies or single innovation projects, decision–oriented measurement practices such as 'benchmarking' or 'scoreboarding' have become well established. Benchmarking is the practice of identifying the organisation (e.g., competitor firm) which is the best at a particular function or activity, such as innovation, and using that organisation's metrics

as the goal to be achieved and surpassed. "Benchmarking is the continuous process of measuring products, services, and practices against the toughest competitors or those companies recognised as industry leaders." (Kearnes, 1986). Although this is clearly a quantitative approach, it also has qualitative aspects. Camp, for example, refers to benchmarking as "the search for industry best practises that lead to superior performance" (Camp, 1989, p. 12). Benchmarking in the industrial context is very action oriented, aimed at improving business operations and competitiveness. The development of indicators and their aggregation are the means to an end in industrial benchmarking.

Let us note here, that, before a decision is made in a firm, the available database can be questioned and even partly laid aside. Here the assessment of quality, although it may be difficult, can be solved in some way. The database with all its biases does not automatically determine decisions. This may be different in the context of a national policy (see below).

Another concept related to S&T indicators that has developed in the business world is that of scoreboarding. Like the sports scoreboards which show how many goals, runs, or points have been scored in a competition or match, scoreboards have been developed and used to show companies how they stand with respect to their competitors on aggregate, widely recognized metrics of business performance such as productivity. In the area of industrial research and development (R&D) and innovation, 'R&D Scoreboards' and even 'Patent Scoreboards' have been developed and published.

Recent years have witnessed the increasing application of these methods, relatively uncritically, for national or regional science and technology policy. In particular, the use of composite indicators is being promoted as an emerging and pioneering field (European Commission 2003 and further references given there on p. 433, in footnote 1). At this level innovation scoreboards and the like are not usually used instrumentally to make policy decisions, because decision making in science and technology policy is quite a complicated negotiation procedure between societal interests and interest groups (Edler et al., 2003). Scoreboards of national innovation performance instead function more as 'soccer league tables' telling the public which countries are performing well or second rate, which have caught up or fallen behind.

The problems with this use of benchmarks or scoreboards on a national level lie in the lack of clear theoretical models that tell us which indicators to select, how to weight them and how to handle cross-country differences in the availability of data (Pohn-Weidinger et al., 2001; European Commission, 2003a). To say the least, this use of scoreboards or benchmarking rank tables may be dangerous because the numbers provided are taken at face value with

little discussion of their validity. Substantial space exists for manipulation by selection, weighting and aggregating indicators. This chapter attempts to raise this point and to provide empirical examples of the range of interpretation or misinterpretation of national innovation scoreboards.

Successful scoreboard–based analysis should depend on mastering the art of indicator selection and scoreboard design. As a *sine qua non* for reasons of public accountability, scoreboards — as any advanced evaluation method — need a clear and transparent structure and recognised concepts (Tijssen, 2003).

3. CALCULATING COMPOSITE INDICES: ONE EXAMPLE

3.1 Methodology

In this chapter we want to investigate the robustness of innovation scoreboards empirically by sensitivity analysis of one selected case. As we have argued above, this seems to be the European speciality driven to a large part by bodies of the European Commission. In any case, it seems to be a newer development within the long standing tradition of S&T indicators. "By aggregating a number of different variables, composite indicators are able to summarise the big picture in relation to a complex issue with many dimensions." (European Commission, 2003, p. 433). What can we learn from aggregating that goes beyond the detailed information?

The procedure will be as follows: We take the original composite indicator of the European Innovation Scoreboard (2001) and compare the ranking of countries by various ranking methods, namely by

- original, Olympic, average and weighted ranks;[2]
- metric scales (weighted and un-weighted) in distinction to rank positions; and
- selective omission of data in order to 'promote' a countries' position.

We determine the weights by grouping of the various indicators into input, throughput, and output and give each group one third weight, but equal weights within the group. There are many metric scales possible; in our case we use the technometric scale originating from evolutionary

[2] The original ranking is explained below, by 'Olympic' we mean that the ranking is done by descending numbers of 'gold medals' first, then by 'silver medals' (rank position 2), and so on.

economics. It adjusts the interval to 'real market' competitive positions (Grupp, 1998).

The empirical base of our case study is the European Innovation Scoreboard 2001 (see section 2.4). In that base, a combination of 18 indicators is presented, namely S&E graduates, tertiary education, lifelong learning, employment in manufacturing and services, R&D intensity, business expenditures on R&D, European and US patents, SMEs' innovation and co-operation, innovation intensity, venture capital, new capital, new products, internet access, information technology markets, and the high tech value added.

From this scoreboard, a 'tentative summary innovation index' (SSI) is constructed, placing, for instance, Sweden in rank position 1 (score 6.5), the UK in rank position 4 (score 4.4), Germany in average rank position 9 (score 0.6) and Greece in rank position 16 (score -7.9) (see Figure 3.1). The SII is equal to the number of indicators which are 20 per cent above average minus the number of indicators that are 20 per cent below. The index is normalised to the interval [10, –10]. An index of zero represents the EU average.

Figure 3.1. Summary innovation index of the European Innovation Scoreboard 2001 (original graph from p. 12)

One may ask several questions, for instance, why this selection of indicators and why is the aggregation done in such a peculiar way? Is it justifiable to give equal weights to 16 out of the 18 indicators but count the two sets of patent data by half each? In addition internal criticism was raised: "While this technique could prove suitable when we have the same

number of indicators across countries, its relevance declines sharply when observations are unevenly distributed across countries, as is the case here." (European Trend Chart, 2003, p. 16, footnote 7). In fact, we face the problem that some data for some countries are missing, which raises special problems. Because we consider benchmarking and score boarding designs an art, we do not want to continue with arguments whether or not precisely this procedure is the best solution of all possible alternatives. But we rather want to process the given data in some other ways and compare the sensitivity of the results (the country ranking) to the original method.

3.2 Robustness of Composite Indices — Selected Results of a Sensitivity Analysis

In Figure 3.2 we provide the results of several aggregation procedures for composite indicators other than the original, as suggested in section 5.[3] What we learn from Figure 3.2 is that Sweden is in rank position 1 irrespective of the aggregation procedure. The same is true for Finland in rank position 2. All the other countries vary by one to four rank positions depending on the aggregation procedure, but overall the impression is that the country ranking cannot be completely turned upside down. The countries in top positions are always in a good position and those at the end of the scale are not positioned in the first half of the league irrespective of the aggregation details.

In particular, the average ranking and the weighted average ranking are most similar to the original SSI index, whereas the Olympic scale can change the picture more seriously. Consider the case of France: France is nowhere in best position amongst the 18 variables, thus wins now a 'gold medal' and will be placed behind all other countries with at least one gold medal (number one in one of the 18 variables).

[3] For EU countries only, thus omitting the US and Japan.

Figure 3.2. Results by selected ranking procedures

Also interesting is the case of the United Kingdom. The UK performs quite strongly in most input indicators but less so in throughput and output variables. As the set of the 18 indicators is very much input biased, equal weighing puts the United Kingdom in a favourable position. When one starts to give all inputs together the same weight as all throughputs and all outputs then this 'natural advantage' in the original SSI ranking vanishes for this country and other countries catch up.

Figure 3.3. Results by selected metric procedures

Figure 3.3 displays the results by comparing the original SII index with metric procedures. Rank tables bear the problem that the distance between any two adjacent positions can be very small in original indicator values or can be large. Metric scales, in distinction, conserve the distances of the original variable values and transform them similarly.

This procedure yields quite different rankings, as is shown in Figure 3.3. Take the example of the United Kingdom again. In the SII scoreboard the country is in third place within the EU countries); however, the distance in most variables to the leading countries is much larger than the rank positions suggests; the United Kingdom seems to be closer to European average if weighted metrics scales are used. Again, the difference is still more pronounced if the weighted metrics are taken, because the 'natural advantage' in input variables vanishes. In the case of Germany the SII index is just above EU average whereas the metric values are clearly below.

Figure 3.4. Some results by selective omission of variables

A third type of sensitivity exercise consists in selective omission of certain variables. We already have argued that scoreboard design is more an art than a science and it is difficult to argue why 18 indicators are used, not 16 or 20, and why exactly these. If we now tune the calculation by selective omission of those variables in which a certain country does not perform very well, we can try to tune the selection of variables in favour of some

countries. In Figure 3.4 we present the results of an optimisation for the UK, for Germany, and for Greece.

The optimisation attempt for the UK is not very successful. The reason is that the set of 18 variables is already optimised in favour of the UK by the original SII (with its many input variables). If we exclude some output and throughput variables in which the UK does not perform very well we cannot really improve the countries' position compared to the original index. This is true for all countries being in lead positions in the original scale; these benefit from the selection by the original scoreboard and cannot really be pushed ahead further.

For countries in middle places and further down the scale one can optimise their position with more success. For instance, Germany can be put in rank position 3 by optimising the selection of indicators, because Germany performs mediocre in life long learning, venture capital, and other variables. If these are taken out of the indicator set the country's position improves considerably. The same is true for Greece, which can be brought upwards by several rank positions if a selection of variables being favourable towards Greece's performance is taken.

How plausible is country tuning? The optimised UK index is achieved by not considering EU patents and giving more weight to US data, not considering business expenditures on R&D but giving more weight to gross expenditures on R&D and to venture capital. These assumptions are not really revolutionary; in particular, in the case of two patent data sets it is questionable why both of them should go into the summary index with half weight (as is the case for the SII). Germany's index, as has been mentioned above, may be improved by not considering venture capital, the opposite assumption as for the UK – but is venture capital really a traditional core S&T indicator? Greece's profile profits from not considering patent indicators at all and leaving out high tech value added.

All these assumptions seem to be soft and can certainly be discussed seriously. Altogether, we think the selection of any one set of indicators does not give equal justice to all countries and thus the selection problem is implicitly a way to tune country positions. This is done by disputable arguments and does not need any heroic assumptions.

4. DISCUSSION AND SUGGESTED RESEARCH QUESTIONS

In this chapter we have given a very brief survey of reports on science and technology indicators in the triad regions. We argue that the European Commission is a latecomer, but is playing a more active role in recent years

and is driven in the direction of composite indicators, which do not seem to be of primary concern in most single country S&T indicator efforts.

By applying alternative aggregation procedures to one selected example (introduced by EU bodies) different weights and a different selection of indicators, we have shown that summary scores and rank positions can vary considerably. It seems that ranking is less sensitive to calculating procedures but gives no information on the size of the gaps. Metric scales seem to provide more insights into relative positions of nations in the very sense of benchmarking. Sensitivity analysis further shows that exclusion or inclusion of variables tend to be a bigger problem than a slight variation of indicator scales. It seems not to be too difficult to argue for a 'country friendly' selection and corresponding weighing of indicators. To say the least, this use of scoreboards or benchmarking tables may be dangerous if the summary numbers provided are taken as such with little discussion of their validity.

The space for manipulation of scoreboards by selection, weighing and aggregation is great. Further research should remedy the situation. This chapter attempts to raise this point and to provide empirical examples of the range of interpretation or misinterpretation of national innovation scoreboards. It did not attempt to suggest more viable alternatives.

Triggering the discussion of these problems, an alternative to scoreboarding of national innovation indices is, nevertheless, suggesting, namely, the use of interval based metric scales in order not to hide the size of the gaps. The use of multi-dimensional representations is the minimum requirement, such as 'spider' charts. Maps of similarity between country structures in science and technology may have more explanatory power in particular when combined with non-quantitative methods.

More research is needed on the validity of S&T indicators, their relationship to important S&T policy concepts, their performance in different science and technology domains, their sensitivity to selection, inclusion, and alternative methods of calculation, as well as their use in the policymaking system and means of preventing their misuse and abuse.

REFERENCES

BMBF (Ed.). (various years). Bundesbericht Forschung. Bonn/Berlin.

Breitzman, A.F., Mogee, M.E. (2002). The many applications of patent analysis. Journal of Information Science, 28 (3), 187–205.

Camp, R.C. (1989). Benchmarking: the search for industry best practices that lead to superior performance. Milwaukee: ASQC Quality Press.

CSRS (Ed.). (1998). Technological innovation in Russia. Moscow: Centre for Science Research and Statistics.

DTI (Ed.). (2003). Competing in the global economy.The innovation challenge. London: DTI.

Edler, J., Kuhlmann, S., Behrens, M. (Eds.). (2003). Changing governance of research and technology policy – The European research area. Cheltenham: Edward Elgar Publishing.
European Commission (Ed.). (1994). The European report on science & technology indicators 1994. Brussels: EUR 15897 EN.
European Commission (Ed.). (2001). European innovation scoreboard 2001. Luxemburg: SEC (2001), 1414.
European Commission (Ed.). (2003). Third European report on science & technology indicators 2003. Brussels: EUR 20025 EN.
European Commission (Ed.). (2003a). Biotechnology innovation scoreboard 2003. Brussels: Innovation/SMEs Programme.
European Commission (Ed.). (2002). Benchmarking national research policies. Brussels: EUR 20494 EN.
European Commission (Ed.). (2003b). Towards a European research area science, technology and innovation – key figures 2003–2004. Brussels: EUR 20735 EN.
Gokhberg, L.N. Gorodnikova and J. Wuttke (1999). Volkswirtschaft im Übergang – Ein Vergleich der FuE-Indikatoren in Russland und Deutschland. Moscow and Essen: CSRS and Stifterverband.
Griliches, Z. (1995). R&D and productivity: econometric results and measurement issues. In Stoneman (Ed.), (pp. 52–89).
Grupp H., Legler, H., Gehrke, B., Breitschopf, B. (2003). Zur technologischen Leistungsfähigkeit Deutschlands 2002. Berlin: BMBF.
Grupp, H. (1998). Foundations of the economics of innovation – theory, measurement and practice. Cheltenham: Edward Elgar Publishing.
Kearnes, D.T. (1986). Quality improvement begins at the top, World, 20 (5), 21.
Kodama, F. (1987). A system approach to science indicators. In Grupp (Ed.), Problems of measuring technological change (pp. 65-87). Köln: Verlag TÜV Rheinland.
Management and Coordination Agency (Ed.). (various years). Report on the survey of research and development, Tokyo.
National Science Board (Ed.). Science (& engineering) indicators, Washington, D.C.
Narin, F., Noma, E. (1985). Is technology becoming science? Scientometrics, 7, 369–381.
Niwa, F., Tomizawa, H. (1995). Integrated indicators: international comparisons of overall strengths in Science and Technology. In: NISTEP (Ed.), Science and Technology Indicators 1994 (pp. 345–365). Tokyo: Science and Technology Agency.
OECD (Ed.). (2003). OECD Science, Technology and Industry Scoreboard. Paris: Organisation for Economic Co-operation and Development.
OST (Ed.). (various years). Science & Technologie – Indicateurs. Paris: Economica.
Patel, P., Pavitt, K. (1995). Patterns of technological activity: their measurement and interpretation. In Stoneman (Ed.), (pp. 15–51).
Pavitt, K. (1988). Uses and abuses of patent statistics. In: A.F.J. van Raan, Handbook of quantitative studies of science and technology (pp. 509–536). Amsterdam: North Holland.
Pohn-Weidinger, S., Gassler, H., Gruber, M., Polt, W., Schibany, A., Leo, H. (2001), Innovationsbericht 2001. Wien: Bundesministerium für Wirtschaft und Arbeit.
Republik Österreich (Ed.). (2003). Österreichischer Forschungs- und Technologiebericht 2003. Wien: Bericht der Bundesregierung an den Nationalrat.
Stoneman, P. (Ed.). (1995). Handbook of the Economics of Innovation and Technological Change. Oxford and Cambridge, Mass.: Blackwell.
Tijssen, R.J.W. (2003). Scoreboards of research excellence. Research Evaluation, 12 (2), 91–103.

Chapter 4

KEEPING THE GATES OF SCIENCE JOURNALS

Gatekeeping Indicators of National Performance in the Sciences

Tibor Braun

Information Science and Scientometrics Research Unit (ISSRU), Institute for Research Organization, Hungarian Academy of Sciences, Budapest, Hungary; and Institute of Inorganic and Analytical Chemistry, L. Eötvös University, Budapest, Hungary
E-mail: braun@helka.iif.hu

Abstract: The chapter outlines the basics of journal gatekeeping and stresses the gatekeeping indicators initiated by us in the 1980s for the evaluation of the research performance of nations in all fields of science. The results have proven that these indicators correlate well with science indicators built on journal paper productivity and citations of nations and represent implicitly a component of quality somewhat associated with citation impact of journal papers. That is why gatekeeping indicators are useful and more simple to build as indicators based on publication productivity and citation impact.

1. INTRODUCTION

The present system of basic research in the sciences and scientific communication depends almost entirely on the primary journal literature. Modern science has developed a particular mechanism of communication which began with the appearance of the first scientific journals in the 17th century and which has remained basically the same ever since. Briefly, this mechanism is based on the selective publication of fragments rather than complete treatises. It is this selective concern with fragments of knowledge, represented primarily by journal articles, that enables science to function effectively and is responsible for its phenomenal growth and pre-eminence (Zuckerman, 1971).

That a paper has been accepted for publication in a well known refereed journal is probably the best immediate indication that it reports worthwhile