# BIBLIOMETRICS AS A RESEARCH FIELD

## A course on theory and application of bibliometric indicators

**W. GLÄNZEL**

COURSE HANDOUTS

2003

**CONTENT**

# GENERAL INTRODUCTION

Bibliometrics has become a standard tool of science policy and research management in the last decades. All significant compilations of science indicators heavily rely on publication and citation statistics and other, more sophisticated bibliometric techniques.

Examples for such compilations are:

- National Science Board
- Observatoire des Sciences et des Techniques
- European Report on S&T Indicators
- Het Nederlands Observatorium van Wetenschap en Technologie: Wetenschaps- en Technologie-Indicatoren
- Vlaams Indicatorenboek

In addition, many extensive bibliometric studies of important science fields appeared during the last two decades. Aim of these studies was to measure national research performance in the international context or to describe the development of a science field with the help of bibliometric means (for instance, *Braun* et al., 1987).

It is a common misbelief that bibliometrics is nothing else but publication and citation based gauging of scientific performance or compiling of cleaned-up bibliographies on research domains extended by citation data. In fact, scientometrics is a multifaceted endeavour encompassing subareas such as structural, dynamic, evaluative and predictive scientometrics. Structural scientometrics came up with results like the re-mapping of the epistemological structure of science based, for instance, on co-citation, ”bibliographic coupling” techniques or co-word techniques. Dynamic scientometrics constructed sophisticated models of scientific growth, obsolescence, citation processes, etc. These models are not only of theoretical interest but can also be usefully applied in evaluation and prediction.

Beyond policy relevant applications of bibliometric results, there are recently important applications in the context of studying the linkage between science and technology, or applications to related fields such as library and information science and most recently also Webometrics. Examples for the latter ones are the large ongoing projects EICSTES (European Indicators, Cyberspace and the Science-Technology- Economy System) and WISER (Web indicators for scientific, technology and innovation research).

Today, bibliometrics is one of the rare truly *interdisciplinary research* fields to extend to almost all scientific fields. Bibliometric methodology comprises components from mathematics, social sciences, natural sciences, engineering and even life sciences. The following pages will provide a systematic description of the research structure of the field and a detailed overview of the state-of-the-art in bibliometric methodology.

# 1. HISTORICAL REMARKS

## *1.1 The origin of the name 'Bibliometrics'*

The terms *bibliometrics* and *scientometrics* were almost simultaneously introduced by *Pritchard* and by *Nalimov* and *Mulchenko* in 1969. While Pritchard explained the term bibliometrics as "*the application of mathematical and statistical methods to books and other media of communication*", *Nalimov* and *Mulchenko* defined scientometrics as "*the application of those quantitative methods which are dealing with the analysis of science viewed as an information process*". According to these interpretations the speciality scientometrics is restricted to the measurement of science communication, whereas bibliometrics is designed to deal with more general information processes. The anyhow fuzzy borderlines between the two specialities almost vanished during the last three decades, and nowadays both terms are used almost as synonyms. Instead, the field *informetrics* took the place of the originally broader speciality bibliometrics. The term informetrics was adopted by VINITI (*Gorkova*, 1988) and stands for a more general subfield of *information science* dealing with mathematical-statistical analysis of communication processes in science. In contrast to the original definition of bibliometrics, informetrics also deals with electronic media and thus includes topics such as the statistical analysis of the (scientific) text and hypertext systems, library circulations, information measures in electronic libraries, models for Information Production Processes and quantitative aspects of information retrieval as well.

## *1.2 The pioneers of bibliometrics*

The statistical analysis of scientific literature began almost 50 years before the term "bibliometrics" was coined. In 1926, *Alfred J. Lotka* published his pioneering study on the frequency distribution of scientific productivity determined from a decennial index (1907-1916) of *Chemical Abstracts.* Lotka concluded that

> *"the number (of authors) making n contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that makes a single contribution, is about 60 per cent."*

This result can be considered as a rule of thumb even today, 75 years after its publication.

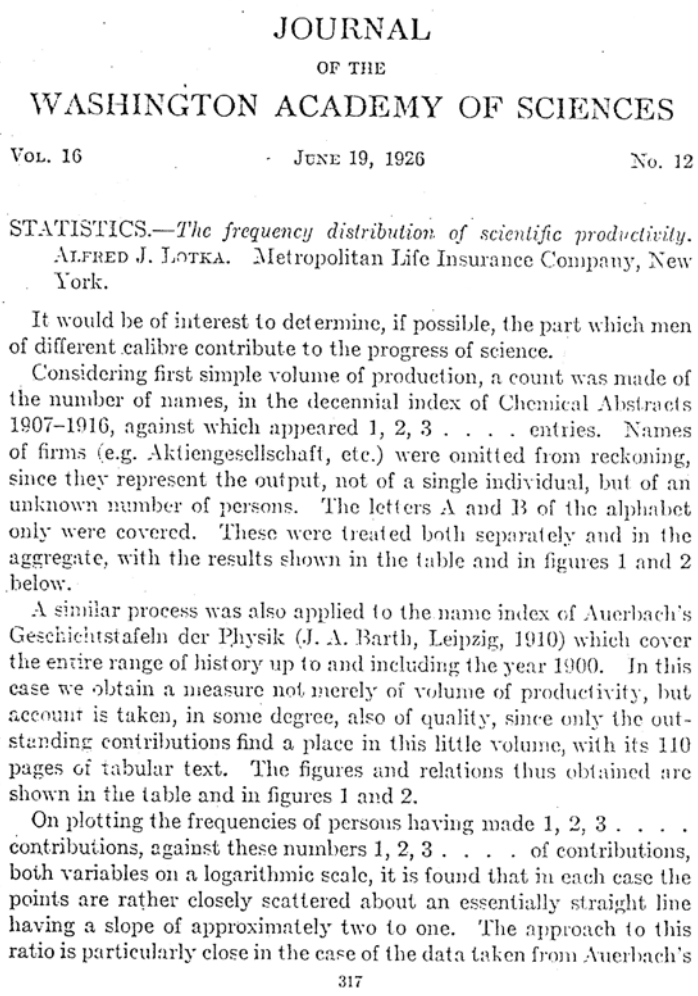
At almost the same time, in 1927, *Gross* and *Gross* published their citation-based study in order to aid the decision which chemistry periodicals should best purchased by small college libraries. In particular, they examined 3633 citations from the 1926 volume of the *Journal of the American Chemical Society*. This study is considered the first citation analysis, although it is not a citation analysis in the sense of present-day bibliometrics.

Eight years after Lotka's article appeared, *Bradford* (1934) published his study on the frequency distribution of papers over journals. He found that

> "*if scientific journals are arranged in order of decreasing productivity on a given subject, they may be divided into a nucleus of journals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus when the numbers of periodicals in the nucleus and the succeeding zones will be as 1: b : $b^2$ ...*"

*Zipf* (1949) formulated an interesting law in bibliometrics and quantitative linguistics that he derived from the study of word frequency in a text. According to Zipf $rf = C$, where $r$ is the rank of a word, $f$ is the frequency of occurrence of the word and $C$ is a constant that depends on the text being analysed. It can be considered a generalisation of the laws by Lotka and Bradford. He formulated the following underlying principle of his law although he has never shown how this principle applies to his equation.

> *"The Principle of Least Effort means… that a person…will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems…."* (*Zipf*, 1949).



JOURNAL
OF THE
WASHINGTON ACADEMY OF SCIENCES

VOL. 16 JUNE 19, 1926 No. 12

STATISTICS.—*The frequency distribution of scientific productivity.* ALFRED J. LOTKA. Metropolitan Life Insurance Company, New York.

It would be of interest to determine, if possible, the part which men of different calibre contribute to the progress of science.

Considering first simple volume of production, a count was made of the number of names, in the decennial index of Chemical Abstracts 1907–1916, against which appeared 1, 2, 3 . . . . entries. Names of firms (e.g. Aktiengesellschaft, etc.) were omitted from reckoning, since they represent the output, not of a single individual, but of an unknown number of persons. The letters A and B of the alphabet only were covered. These were treated both separately and in the aggregate, with the results shown in the table and in figures 1 and 2 below.

A similar process was also applied to the name index of Auerbach's Geschichtstafeln der Physik (J. A. Barth, Leipzig, 1910) which cover the entire range of history up to and including the year 1900. In this case we obtain a measure not merely of volume of productivity, but account is taken, in some degree, also of quality, since only the outstanding contributions find a place in this little volume, with its 110 pages of tabular text. The figures and relations thus obtained are shown in the table and in figures 1 and 2.

On plotting the frequencies of persons having made 1, 2, 3 . . . . contributions, against these numbers 1, 2, 3 . . . . of contributions, both variables on a logarithmic scale, it is found that in each case the points are rather closely scattered about an essentially straight line having a slope of approximately two to one. The approach to this ratio is particularly close in the case of the data taken from Auerbach's

317

*Figure 1 First page of Alfred Lotka's famous article on scientific productivity of chemists*

Relatively little attention has been paid to these results. The causes for this phenomenon are threefold.

1. These papers appeared when traditional methods of information retrieval were still sufficient,
2. they applied to different phenomena and the interrelation between these laws which was not completely recognised;

3. and financing systems for scientific research did not yet stand need of quantitative or even sophisticated statistical methods.

This situation dramatically changed when *Derek deSolla Price* published his fundamental work in bibliometrics.

### *1.3 Bibliometrics since deSolla Price*

In his book entitled "Little Science – Big Science" (1963), *Derek deSolla Price* analysed the recent system of science communication and thus presented the first systematic approach to the structure of modern science applied to the science as a whole.

At the same time, he laid the foundation of modern research evaluation techniques. *DeSolla Price*' work was more than pioneering; it was revolutionary. Time was now ripe for the reception of his ideas since globalisation of science communication, the growth of knowledge and published results, increasing specialisation as well as growing importance of interdisciplinarity in scientific research reached a stage where scientific information retrieval began to fail and funding systems based on personal knowledge and evaluations by peer reviews became more and more difficult.

At that time, most basic models for scientific communication were developed. Among these are first models for essential concepts in scientific communication like growth and ageing of information. Literature and information was assumed to grow exponentially, but in individual research disciplines the growth can also be linear or logistic. Finally, the logistic model has been widely accepted since both exponential and linear growth can be considered special phases within the logistic model. The concept of ageing or obsolescence is intimately linked with the growth of science. In information science and bibliometrics, changing frequency of citations given or received over time is assumed to reflect ageing of scientific literature. Some authors have downright considered growth and obsolescence inverse functions, the faster growth of literature in a field, the faster it ages and the literature becomes obsolete in a shorter time (*Brookes*, 1970*, Egghe*, 1993, *Kärki* and *Kortelainen*, 1998). Consequently, an exponential model has been proposed for the ageing of literature in the beginning, too (*Wallace*, 1986, *Price*, 1963). In particular, the model of *radioactive decay* has been adopted. Later on, more complex models have been developed (see, *Glänzel* and *Schoepflin*, 1995, 1999, *Egghe*, 1993).

*Goffman* and *Nevill* have introduced the theory of *intellectual epidemics as a model of scientific communication* in 1964. According to this model the diffusion of ideas in a population of scientists could be compared to the spreading of an influenza virus in a population of people, causing an epidemic. This model can be used both, to describe the spread of the disease and to predict the time when the disease reaches its peak, after which it is presumed to decline. The advantage of the model lies in its predictive power. *Goffman* and *Nevill* proposed that the same model could also describe the spread of information within the scientific community. According to the model, the population can at any time be subdivided into three groups of infected, resistant and infection sensitive persons. If a published article in a specific topic is considered an infection, it is possible to follow the diffusion of the epidemic by counting the number of publications per author and theoretically make a forecast of its future. Communication between authors builds on attempts to distribute ideas aiming at reception of disseminated information and on providing contact between infection susceptible and already infected persons.

Another general theory characterising processes of scientific communication is the principle of cumulative advantage. *Price* formulated this in 1976 as follows.

> *"Success seems to breed success. A paper which has been cited many times is more likely to be cited again than one which has been little cited. An author of many papers is more likely to publish again than one who has been less prolific. A journal which has been frequently consulted for some purpose is more likely to be turned to again than one of previously infrequent use."*

Bibliometrics/Scientometrics took a sharp rise since the late sixties. In the seventies, when data collection was often still a matter of manual work, the field *bibliometrics* was, characterised by the personalities of enthusiastic researchers much in the way of a "hobby" to later integrate interdisciplinary approaches as well as mathematical and physical models on one side, and sociological and psychological methods on the other, not speaking of the long tradition of library science. Later on, since the beginning of the eighties, bibliometrics could evolve into a distinct scientific discipline with a specific research profile, several subfields and the corresponding scientific communication structures (publication of the international journal *Scientometrics* in 1979 as the first periodical specialised on bibliometric topics, international conferences since 1983, the journal *Research Evaluation* since 1991).
The main reason for this development can be seen in the availability of large bibliographic databases in machine-readable form and the fast development of computer science and technology. This made it possible that metrics of science could be established also outside the USA. First, license fees and expensive CPU time resulted at least in the 80s in severe limitations but the technology of the 90s brought the breakthrough. "On-line bibliometrics", however, remained a dream.
The funding of large projects seems to have become the regular way of financing research in scientometrics. From "Little Scientometrics" the field has become "Big Scientometrics". The publication of several comprehensive books on bibliometrics, among others by Haitun (1983), Ravichandra Rao (1983), Bujdosó (1986), van Raan (1988), Egghe and Rousseau (1990), and Courtial (1990), may reflect this process. The fact that bibliometric methods are already applied to the *field* "bibliometrics" itself also indicates the rapid development of the discipline.

### *1.4 The three "components" of present-day bibliometrics*

Present-day bibliometric research is aimed at the following three main target-groups that clearly determine topics and sub-areas of "contemporary bibliometrics".

(i) *Bibliometrics for bibliometricians (Methodology)*

This is the domain of basic bibliometric research and is traditionally funded by the usual grants. Methodological research is conducted mainly in this domain.

(ii) *Bibliometrics for scientific disciplines (Scientific information)*

The researchers in scientific disciplines form the bigger, but also the most diverse interest-group in bibliometrics. Due to their primary scientific orientation, their interests are strongly related to their speciality. This domain may be considered an extension of *science*

*information* by metric means. Here we also find joint borderland with quantitative research in *information retrieval.*

(iii) *Bibliometrics for science policy and management (science policy)*

This is the domain of *research evaluation*, at present the most important topic in the field. Here the national, regional, and institutional structures of science and their comparative presentation are in the foreground.

Finally, we will have a look at how bibliometrics/scientometrics is linked with related fields and application services (see Figure 1.1).

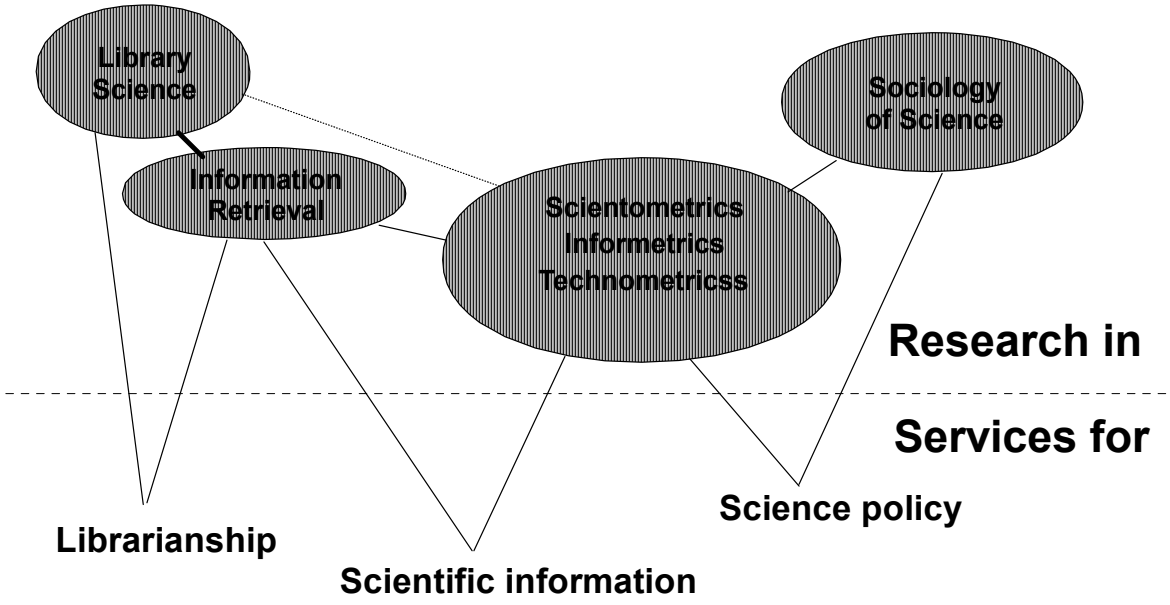



*Figure 1.1 Links of bibliometrics with related fields and application services*

## 2. THE ELEMENTS OF BIBLIOMETRIC RESEARCH AND THEIR MATHEMATICAL BACKGROUND

According to Pritchard (1969), bibliometrics is *the application of mathematical and statistical methods to books and other media of communication.* This basically comprises books, monographs, reports, theses, papers in serials and periodicals, and nowadays also e-books and e-journals as well as – in the broadest sense – the WEB. Nevertheless, periodicals have played the most important part in communication in the sciences. The *Philosophical Transactions of the Royal Society* and the *Journal des Sçavans*, that both appeared first in the year 1665, are considered the first scientific journals.

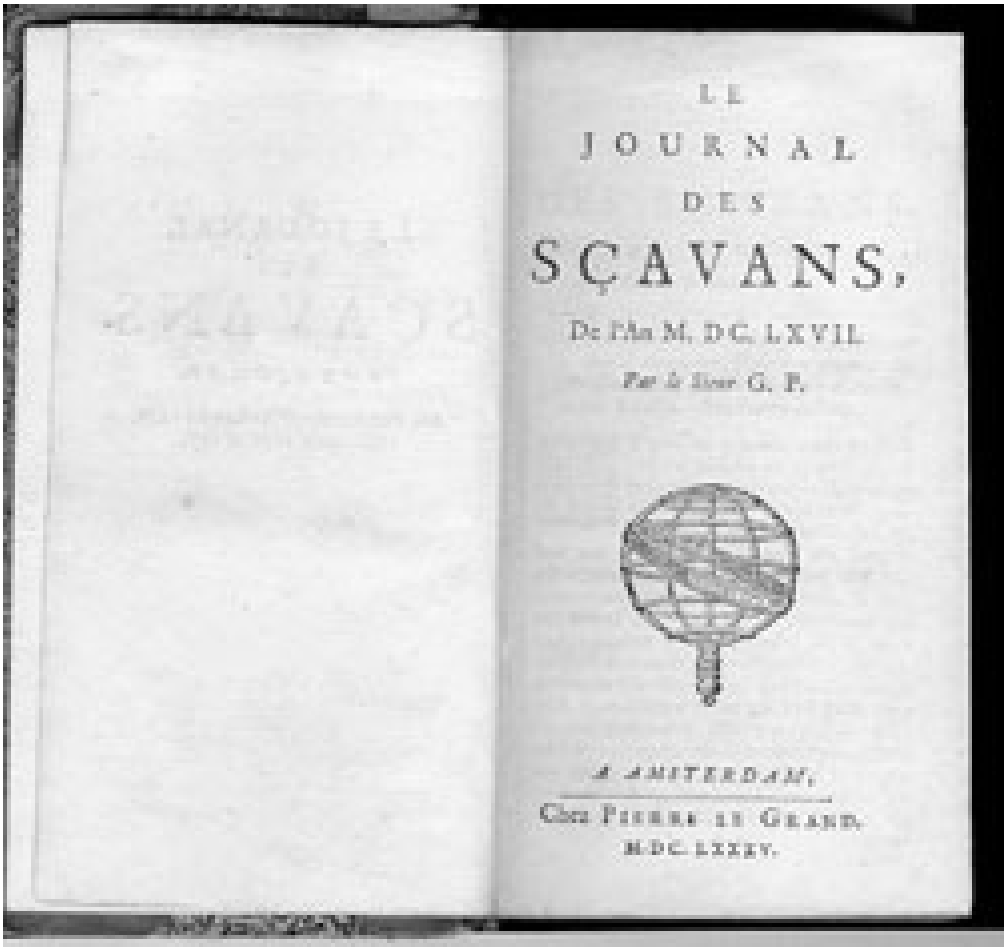

LE
JOURNAL
DES
SÇAVANS,
De l'An M. DC. LXVII.
Par le Sieur G. P.
A AMSTERDAM,
Chez PIERRE LE GRAND.
M. DC. LXXXV.

*Figure 2.1 Cover page of the Journal des Sçavans*

The following figure visualises the growth of the number of scientific journals and review journals since 1665 (according to Price, 1963).

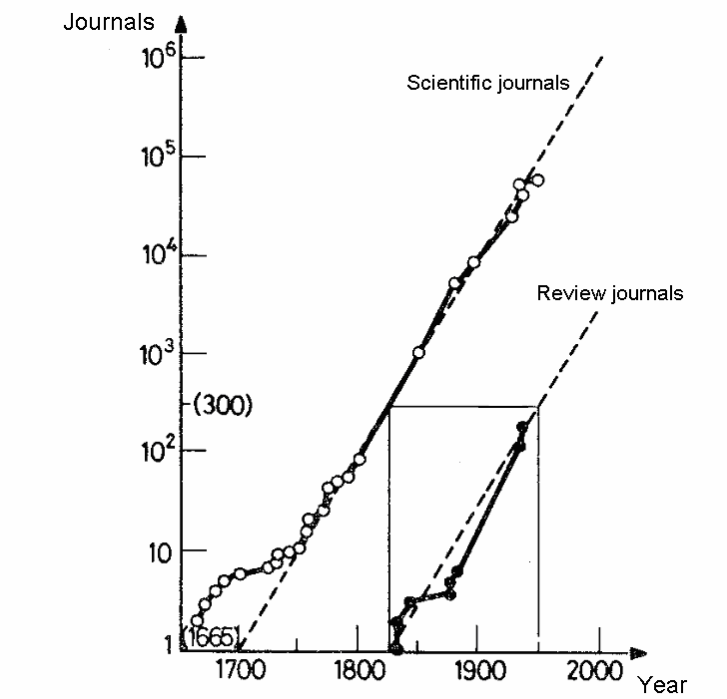



*Figure 2.2 Growth of the number of periodicals (according to Price)*

## *2.1 Basic concepts of elements, units and measures of bibliometric research*

As mentioned above, books, monographs, reports, theses and papers in serials and periodicals are units of bibliometric analyses. Since certain standards are postulated for such units, the scientific paper published in refereed scientific journals proved to be the unit most suitable for bibliometric studies. Among the common standards, we find the reviewing system, the criterion of originality of research results, the availability of literature and the more or less transparent rules. The scientific paper has become the basic unit of bibliometric research. Although structural analyses of the scientific paper are conducted (for example, *Mullins* et al., 1988), there are basic objects like the *paper* that are usually not further subdivided. These form the *elements* of bibliometric analyses. Further elements besides *publications*, are *(co-)authors*, *references* and *citations*.
Publications can be assigned to the journals in which they appeared, through the corporate addresses of their authors to institutions or countries, references and citations to subject categories, and so on. Under given conditions, specific sets of elements can thus be defined; these form *units*. Such units are the already mentioned *journals*, *subject categories*, *institutions* and *countries* to which elements can – not necessarily uniquely – be assigned. Instead of this somewhat lax verbal description, a more precise definition within the framework of appropriate mathematical models is possible, although this is hardly used in the relevant literature. Nevertheless, we will sketch a mathematical model in a later section.

*Basic measures* are simple counts such as publication counts or the number of co-authors or the number of citations received by a set of publications or of the number of given bibliometric units. From the mathematical viewpoint, these measures can be represented by "natural counting measures", namely, the cardinality of the intersection or union of bibliometric units. Figure 2.1 may just serve as an example for a basic measure, namely for the annual change of the number of journals.

More complex measures can be obtained as statistical functions defined on sets of bibliometric elements and units. These measures are also called *bibliometric indicators.* The fundamental demand upon bibliometric indicators is their *validity*, that is, we have to make sure that we really measure what we are intending and assuming to measure. Also *reproducibility* is one of the basic criteria in scientific research. Under identical conditions research results should be reproducible in bibliometrics, too. The reproducibility of results can only be guaranteed, if all sources, procedures and techniques are *reliable* and properly *documented* in scientific publications.

Elements, units and measures of bibliometric research will be discussed in detail in the following sections devoted to concrete applications.

### *2.2 Data sources of bibliometric research*

Data sources of bibliometric/scientometric research and technology are bibliographies and bibliographic databases. Bibliometric analyses can be conducted on the bases of any sufficiently large publication list compiled and issued, for instance, by a scientific institution. Nevertheless, most reliable sources are the big specialised or multidisciplinary databases that have been first provided in printed form but later on also in electronic form (magnetic tape, CD-ROM, on-line). Prominent specialised databases are among others *Medline* (life sciences), *Chemical Abstracts* (chemistry-related literature and patents(!)), *Inspec* (physical sciences and engineering) and *Mathematical Reviews (*mathematics).

The databases of the Institute for Scientific Information (ISI, Philadelphia, PA, USA), first of all, the *Science Citation Index* (SCI) have become the most generally accepted basic source for bibliometric analyses. Although, there are several objections against the journal coverage and the data processing policy of the ISI in preparing the SCI, its unique features are basic requirements of bibliometric technology. Among these features we mention

- *Multidisciplinarity*. All research fields in the life sciences, natural sciences, mathematics and engineering are represented.
- *Selectiveness*. Periodicals covered by SCI are chosen on the basis of quantitative criteria (impact measurements), and the selection is generally reinforced by expert opinion.
- *Full coverage*. All papers published in periodicals covered by the SCI are recorded.
- *Completeness of addresses*. The addresses of all authors are indicated, allowing analyses of scientific collaboration and the application of full publication counting schemes.
- *Bibliographical references.* Together with each document their references are processed. Redefining references as sources makes it possible to analyse citation patterns and to construct citation indicators.
- *Availability*. The SCI is available as printed edition, in electronic form on magnetic tapes, as on-line version and as CD-ROM edition. Especially the latter one gained popularity in the nineties. The Web of Science, however, is rather a tool for traditional information retrieval tasks.

The SCI and the SSCI (*Social Sciences Citation Index*) databases also have – from the viewpoint of bibliometric application – shortcomings. Being multidisciplinary databases they do not provide a direct subject assignment for recorded papers. For lack of an appropriate subject-heading system, versions of ISI's *Subject Category* scheme published in annual *Science Citations Index Guides* and the *ISI Journal Citation Reports* (*JCR*) are used for bibliometric application as an indirect subject assignment of individual papers through those journals in which they have been published. Such assignment system based on journal classification has been developed by *Narin* and *Pinski* (see, for instance, *Narin*, 1976). Since journals are often not devoted to a single topic, assignment of subject areas through journal classification is necessarily less precise than that based on subject headings of individual publications.

*Data base availability (SCI, WoS)*

The SCI database was available already in the '70s. The printed version can still be found in many libraries. The main components of this edition were the following three indexes:

- Source Index
- Citation Index
- Permutation Index

Each covers the same material but indexes it differently. There is a large range of search options the SCI and SSCI offer. The literature is indexed in different ways:

- by cited author and cited work or by cited patent (Citation Index)
- by source author (Source Index) or by source organization (Corporate Index, a section of the source index)
- by title words (Permuterm Subject Index)

The Citation Index is the main search tool; the other indexes, however, play important complementary roles. The Source Index can be used for author searches, and provides a full bibliographic description of every indexed item. The Corporate Index permits searches by organisations.
The Permuterm Subject Index (PSI) goes far beyond a conventional title-word index. The PSI lists under each term all the other title words with which it has appeared. The refinement enables the user to search a combination of two terms, thus increasing the specificity of the search and decreasing the percentage of irrelevant material.
Despite of the unique features of this database, bibliometric research on the basis of the printed edition was restricted to quite small samples. However, the same material was provided on magnetic tapes for retrieval purposes and for scientific information. This information service was based on weekly updates. ISI provided the annual cumulations as an additional service for a supplementary fee. These cumulations could already be used for computerised data processing within the framework of bibliometric studies. Data processing on mainframe computers was, however, limited by storage, speed and above all, by the expensive CPU-time.

```
FN ISI Export Format
VR 1.0
PT Journal
AU Kostoff, RN
   Braun, T
   Schubert, A
   Toothman, DR
   Humenik, JA
TI Fullerene data mining using bibliometrics and database
   tomography
SO JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES
LA English
DT Article
NR 19
SN 0095-2338
PU AMER CHEMICAL SOC
C1 Off Naval Res, 800 N Quincy St, Arlington, VA 22217 USA
   Off Naval Res, Arlington, VA 22217 USA
   Lorand Eotvos Univ, Inst Inorgan & Analyt Chem, H-1443 Budapest, Hungary
   Hungarian Acad Sci Lib, Budapest, Hungary
   RSIS Inc, Mclean, VA USA
   Noesis Inc, Arlington, VA 22203 USA
AB Database tomography (DT) is a textual database analysis system
   consisting of two major components: (1) algorithms for
   extracting multiword phrase frequencies and phrase proximities
   (physical closeness of the multiword technical phrases) from
   any type of large textual database, to augment (2)
   interpretative capabilities of the expert human analyst. DT was
   used to derive technical intelligence from a fullerenes
   database derived from the Science Citation Index and the
   Engineering Compendex. Phrase frequency analysis by the
   technical domain experts provided the pervasive technical
   themes of the fullerenes database, and phrase proximity
   analysis provided the relationships among the pervasive
   technical themes. Bibliometric analysis of the fullerenes
   literature supplemented the DT results with
   author/journal/institution publication and citation data.
   Comparisons of fullerenes results with past analyses of
   similarly structured near-earth space, chemistry,
   hypersonic/supersonic flow, aircraft, and ship hydrodynamics
   databases are made. One important finding is that many of the
   normalized bibliometric distribution functions are extremely
   consistent across these diverse technical domains and could
   reasonably be expected to apply to broader chemical topics than
   fullerenes that span multiple structural classes. Finally,
   lessons learned about integrating the technical domain experts
   with the data mining tools are presented.
CR *ENG INF INC, 1999, ENG COMP
   *I SCI INF, 1999, SCI CIT IND
   ANWAR MA, 1997, SCIENTOMETRICS, V40, P1
   BRADFORD SC, 1934, ENGINEERING, P137
   GARFIELD E, 1985, J CHEM INF COMP SCI, V25, P170
   KOSTOFF RN, 1995, 5440481, US
   KOSTOFF RN, 1997, ADA296021 DTIC
   KOSTOFF RN, 1994, COMPETITIVE INTELLIG, V5, P1
   KOSTOFF RN, 1993, COMPETITIVE INTELLIG, V4, P1
   KOSTOFF RN, 1998, INFORM PROCESS MANAG, V34, P69
   KOSTOFF RN, 1999, J AM SOC INF SCI
   KOSTOFF RN, 1997, J INFORM SCI, V23, P301
   KOSTOFF RN, 1997, SCI ENG ETHICS, V3, P2
   KOSTOFF RN, 1998, SCIENTOMETRICS, V43, P27
   KOSTOFF RN, 1997, SCIENTOMETRICS, V40, P103
   KOSTOFF RN, 1997, SCIENTOMETRICS, V39, P225
   KOSTOFF RN, 1999, TECHNOVATION, V19, P593
   LOTKA AJ, 1926, J WASH ACAD SCI, P16
   MACROBERTS M, 1996, SCIENTOMETRICS, V36, P3
TC 11
BP 19
EP 39
PG 21
JI J. Chem. Inf. Comput. Sci.
PY 2000
PD JAN-FEB
VL 40
IS 1
GA 278YP
PI WASHINGTON
RP Kostoff RN
   Off Naval Res, 800 N Quincy St, Arlington, VA 22217 USA
J9 J CHEM INFORM COMPUT SCI
PA 1155 16TH ST, NW, WASHINGTON, DC 20036 USA
UT ISI:000085016800002
ER
```

*Figure 2.3 Complete bibliographic information about a paper by Kostoff et al. (2000) according to the SCI Expanded*

The on-line versions of the SCI and SSCI databases, for instance, SCISEARCH and SOCIAL SCISEARCH offered by DIMDI (Cologne), seemed to be a serious alternative to the off-line editions also for bibliometric research ("on-line bibliometrics". Although, DIMDI offered many pre-bibliometric options, prices, speed and limited options were the main causes why on-line bibliometrics remained a dream.

In the 90s, especially the CD-Editions of the three ISI databases, the SCI, SSCI and the AHCI (*Arts & Humanities Citation Index*), have become very popular. Besides the standard edition, ISI also provided a version with abstracts. The basic index can be considered an extension of the PSI. From the second half of the 90s on, the excellent retrieval system provided with the CD database permitted bibliometric work on any advanced PC system. Nevertheless, all data downloaded or extracted from the CD-Edition need a careful cleaning-up. The drawback of this popularity was the appearance of bibliometric results based on "quick-and-dirty" searches.
The *Web of Science* is an online edition that combines the three ISI databases SCI expanded (an SCI edition with broader coverage), the SSCI and the ACHI in a unique on-line database. The SCIE covers about 5900 journals whereas the SCI covers about 3500, the SSCI covers 1700 journals and 3300 journals selectively, the ACHI finally covers more than 1100 journals fully and about 7000 journals selectively. The WoS, in turn, is part of the more comprehensive *Web of Knowledge*. The WoK comprises the above-mentioned ISI databases as well as the *Derwent Innovations Index*, *BIOSIS Previews*, *ISI Proceedings*, *CAB ABSTRACTS* and *INSPEC* bibliographic and patent databases. Figure 2.3 gives an example for the bibliographic information provided by the WoS.

### *2.3 Minimum bibliographic description for paper identification*

Bibliometric analyses are based on relevant information about scientific publications that can be retrieved from the above-mentioned data sources. In the following table, most important information used in bibliometrics is shown. This information is usually organised in corresponding search fields.

*Relevant information from bibliographic databases:*

*Scientific publications*

1. Source identification (Journal title, PY, VOL, 1st page)
2. Names of authors
3. Corporate addresses
4. References
5. Document type
6. Title, controlled terms, keywords, abstract, subject headings
7. Acknowledgement

Although no systematic research has been done in this topic, the determination of the minimum description of bibliographic items plays in important part in bibliometric research and technology. This is necessary by several reasons, namely, because possible internal identification numbers cannot be used outside the used database, to correct errors (spelling

variances) in database fields, to be able to match items taken from different fields (reference and source items for citation analyses) and to unify transcription standards in different databases when information from several bibliographic databases is combined. The following example might illustrate the necessity for finding a unified bibliographic description for paper identification.

*Example (reference items)*

> **ZWIE**TERING- M -19**91**-**A**PPL-ENVIRON-MICROB-V**57**-P**1094**
> **ZWIE**TERING- MH -19**91**- **A**PPL-ENVIRON-MICROB -V**57**-P**1094**
> **ZWIE**TERING-MH-19**91**- **A**PPLIED-ENV-MICROBIO -V**57**-P**1094**
>
> *Source: Jan-Sep 1993 cumulation of the 'Science Citation Index' (CD-Edition)*

Three different variants have been found for one and the same paper in the reference strings. According to empirical investigations, the above errors can be corrected, for instance, with the help of the *cluster-key* '**91ZWIE 571094[A]**'.

## *2.4 Mathematical models and the "distributional" approach*

*Pritchard* (1969) explained the term bibliometrics as "the application of mathematical and statistical methods to books and other media of communication". The creation and application of mathematical models seems in this context quite obvious. Moreover, the links created by co-authorship relations, by received and given citations form complex networks in scientific communication that can best be described and analysed with the help of mathematical tools.

To sum up, the application of mathematical models, especially that of stochastic models and probability distributions has several advantages, in particularly, it helps to give

1. mathematical interpretations beside the bibliometric ones,
2. understanding of complex structures such as communication networks,
3. information about statistical reliability and estimates of random errors.

### *2.4.1 Historical sketch of mathematical methods used in bibliometrics*

As in most young emerging fields, also bibliometricians have first adopted models from other fields to describe basic regularities they have observed. And as in most such cases, these models proved to have limited validity. Lotka's Law according to which authors are producing $n_1/n^2$ papers, where $n_1$ is the number of authors having published only one article and $n$ the number of papers, is a typical example. Lotka has formulated as a *natural law*, not taking into account important factors heavily influencing productivity. These factors will be studies later.

A second law was formulated in the context of the growth of literature. The chart presented in Figure 2.2 reflects a more or less *exponential growth* of the number of scientific journals and of review journals. This has been generalised a general law characterising the growth of scientific literature and of information, in general. A consequence, the growth rate is assumed

to be linear at any time. At lower levels of aggregation, for instance, in individual research disciplines the growth has often be assumed to be logistic.

*The exponential model of growth*

In this model, the (cumulative) number of scientific publications $p(t)$ is considered a function of time. Moreover, the growth rate is assumed to be linear, that is, $p'(t) = k \cdot p(t)$, where $k$ is a positive real value. This simple differential equation with the initial condition $p(0) = p_0$ leads directly to the exponential function $p(t) = p_0 \cdot e^{k \cdot t}$. This growth curve can be characterised by the time $t_d$ during which the value of $p_0$ has doubled. In particular, we have $t_d = (\log 2)/k$. Consequently, $t_d$ can be used as a substitute of the parameter $k$ as follows: $p(t) = p_0 \cdot e^{(\log 2) \cdot t/t_d}$.

*The logistic growth model*

It is obvious that the growth of literature in a given topic cannot be exponential till infinity. Beyond a certain threshold, the gradient of the linear growth function begins to decrease and growth turns negative as "resources are exhausted". System finally converges to a level of saturation. This model is described by the *logistic curve* suggested by Pearl-Reed that can be obtained as a solution of the following non-linear differential equation:

$p'(t) = k \cdot p(t)(b - p(t))$; $p \in (0, b)$, where $k$ and $b$ are a positive real values and $p \in (0, b)$.

Under the initial condition $p(0) = p_0$, we obtain the following solution.

$$p(t) = b \cdot [1 + (b/p_0 - 1) \cdot e^{-kbt}]^{-1}, \qquad k > 0.$$

The following figure visualises the logistic curve that consists of three "parts", a quasi-exponential growth if $p(t) \ll b$, a quasi-linear growth if $p(t) \sim b/2$ and finally a negative exponential growth if $p(t)$ is close to level of saturation $b$.

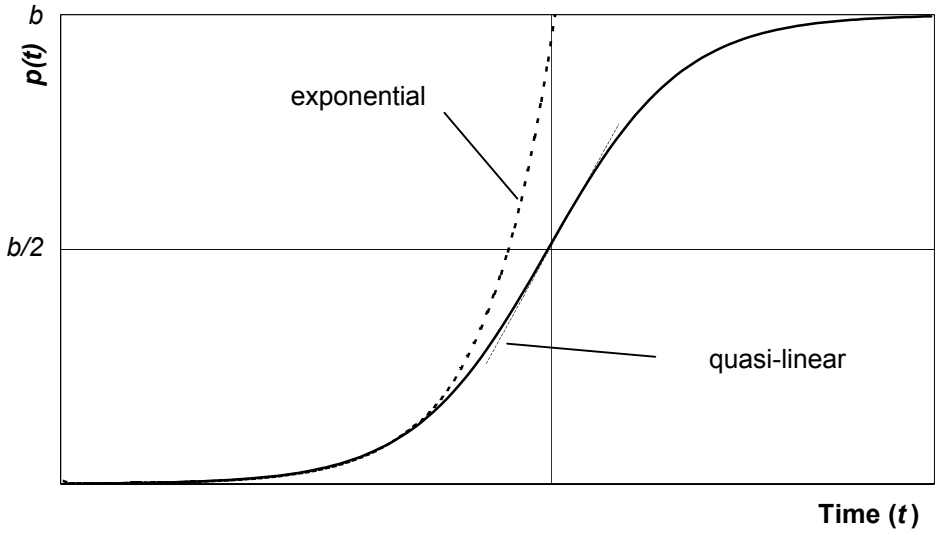



*Figure 2.4 The logistic curve*

*The linear growth model*

We just mention in passing that a linear model can approximate the growth in shorter periods. This applied above all for the phase around the turning point in the logistic model. The following example presents results from a "quick-and-dirty" retrieval based on the keyword *magainin*. Magainin has found in the skin of certain frog species, and belongs to a broad class of antimicrobial peptides that kill bacteria by permeabilising the cytoplasmic membrane. Unlike conventional antibiotics, magainin does not cause resistance. The results of the literature search in the WoS reflect almost linear growth for the 5-year period 1998-2002 with an annual average growth rate of 41 papers (see Figure 2.5). Nevertheless, the growth rate proved to be roughly linear (see Figure 2.6). This reflects a slow exponential growth typical of a small emerging field. The evolution practically started from zero in 1987 and, from 1988 on, growth tripled in annual rate continuously.

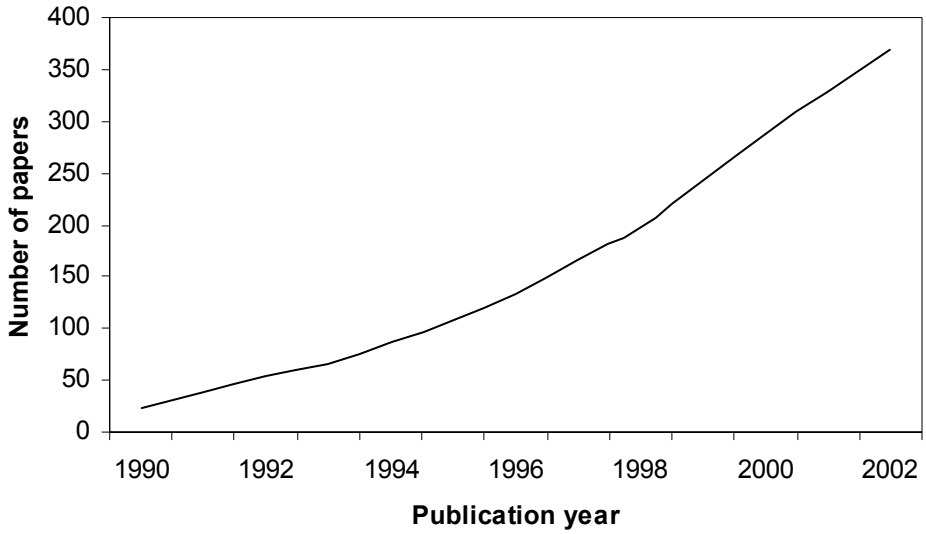



*Figure 2.5 The approximately linear growth of "magainin research" in the last five years*

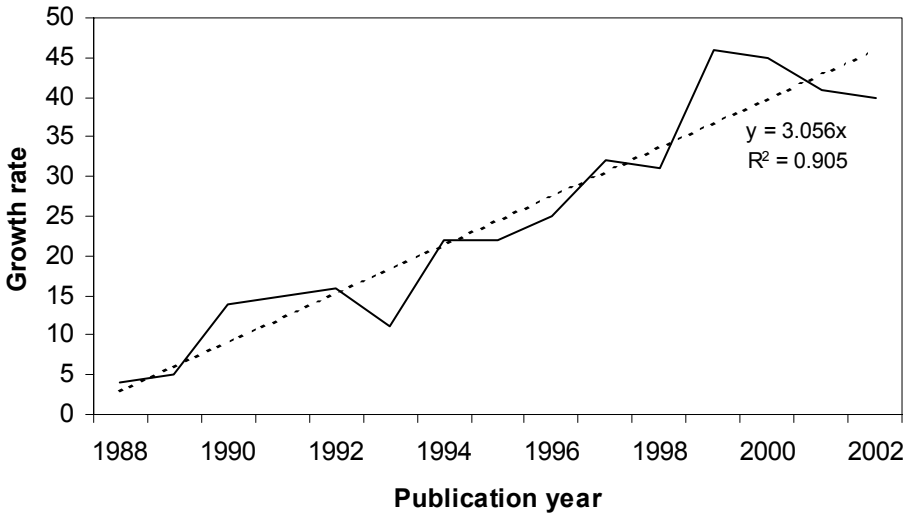



*Figure 2.6 The linear annual growth rate of "magainin research"*

*Intellectual epidemics as a model of scientific communication*

*Goffman* and *Nevill* have formulated their deterministic model of *intellectual epidemics* in 1964. Their model is based on the classical Reed-Frost model. According to this model the diffusion of ideas in a population of scientists could be compared to the spreading of an influenza virus in a population of people, causing an epidemic. The population can at any time be subdivided into three groups of infected ($I$), resistant or immune ($R$) and infection sensitive, i.e., susceptible ($S$) animals or persons. *Goffman* and *Nevill* have considered a published article in a specific topic an infection.
The *deterministic model* proceeds from the classical Reed-Frost model, and can be described by the following system of differential equation. First, we assume that the population is closed, that is,

$S(t) + I(t) + R(t) = N$, where $N$ is positive real constant and $t$ is the time parameter.

For the system of differential equation, we assume:

(1) $S'(t) = -\beta S(t) \cdot I(t)$

(2) $I'(t) = \beta S(t) \cdot I(t) - \gamma I(t)$

(3) $R'(t) = \gamma I(t)$

In order to stabilise the epidemic model, we have to assume that

(2) $I'(t) = \beta S(t) \cdot I(t) - \gamma I(t) > 0$ , that is, $S(t) > \gamma/\beta$.

The epidemic situation occurs at a time $t_0$ when the number of susceptible persons ($S_0$) exceeds the ratio $\gamma/\beta$. We obtain the peak of the epidemic by determining the following extreme value. The change of the number of infected (I) and susceptible (S) persons in time takes its maximum if $[I'(t) + S'(t)]' = [-\beta S(t) \cdot I(t) + \beta S(t) \cdot I(t) - \gamma I(t)]' = [-\gamma I(t)]' = 0$. Since $\gamma \neq 0$, the right-hand side of Eq (2) must vanish, i.e., $S(t) = \gamma/\beta$. This contradicts the condition that the epidemic situation occurs if $S$ exceeds this value. Consequently, we have to assume an open population, that is, we have to assume that $N$ is a increasing function of time. According to Goffman, an open population has to be assumed since sensitive and infected persons have to be continuously replaced by persons entering the system. Here, we just refer to a model by *Schubert* and *Glänzel* introduced in 1983 to describe stationary publication processes. This model will be discussed somewhat later, in the context of stochastic processes.
In the case of the open population, the condition for reaching the peak can be derived analogously to the previous case. In particular, we obtain $S(t) + I(t) = \text{const}$ or $R'(t) = \text{const}$ as sufficient conditions for reaching the peak.

### *2.4.2 Basic postulates and the "axiomatic approach" to bibliometrics*

We can formulate the following basic postulates or even axioms without which bibliometric measures cannot properly be defined and used.

1. A paper receives at one time at most one citation.
2. An author publishes only one article at one time.
3. A paper does not receive citations prior to it its publication
4. The citation link between two papers is unique

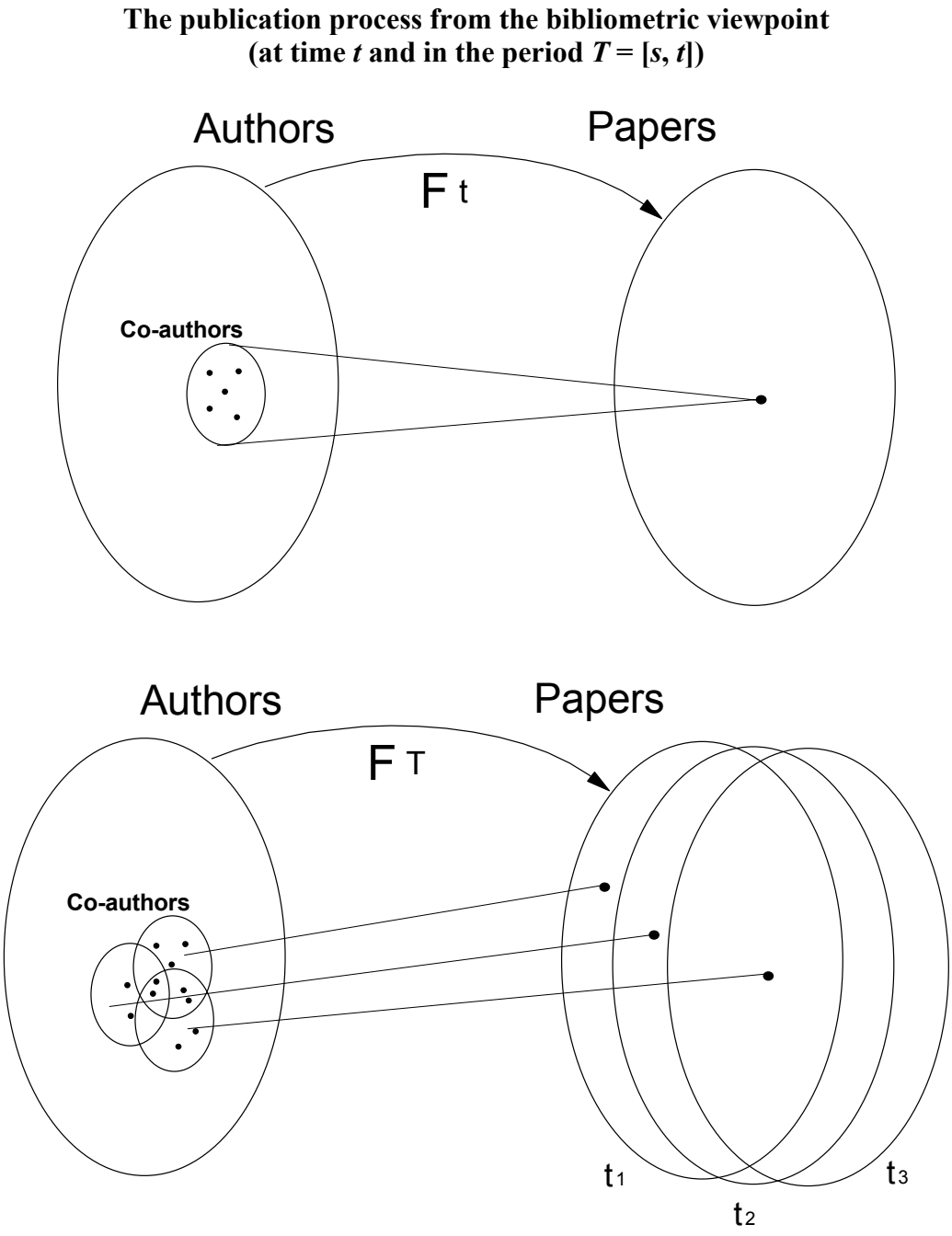



*Figure 2.7 The relationship between bibliometric elements and units in terms of their mathematical interpretation (author–publication links)*

The first two postulates are necessary to define mappings and allow the application of point processes. Sometimes occurs, however, that a paper receives more than one citation in the same issue of a journal, that is, by papers that appear at the same time. This can be solved by using an infinitesimal number $\varepsilon > 0$, so that the paper is, for instant, cited by the publications of the same issue at time $t$, $t + \varepsilon$, $t + 2\varepsilon$, etc. The case that an author has more than one paper in one and the same issue of a journal can be treated analogously.
The second postulate is straightforward and the third one is necessary to make quantification of citation impact possible. The database producer normally guarantees uniqueness of citation links.

**The citation process from the bibliometric viewpoint**
**(at time *t* and in the period *T* = [*s*, *t*])**

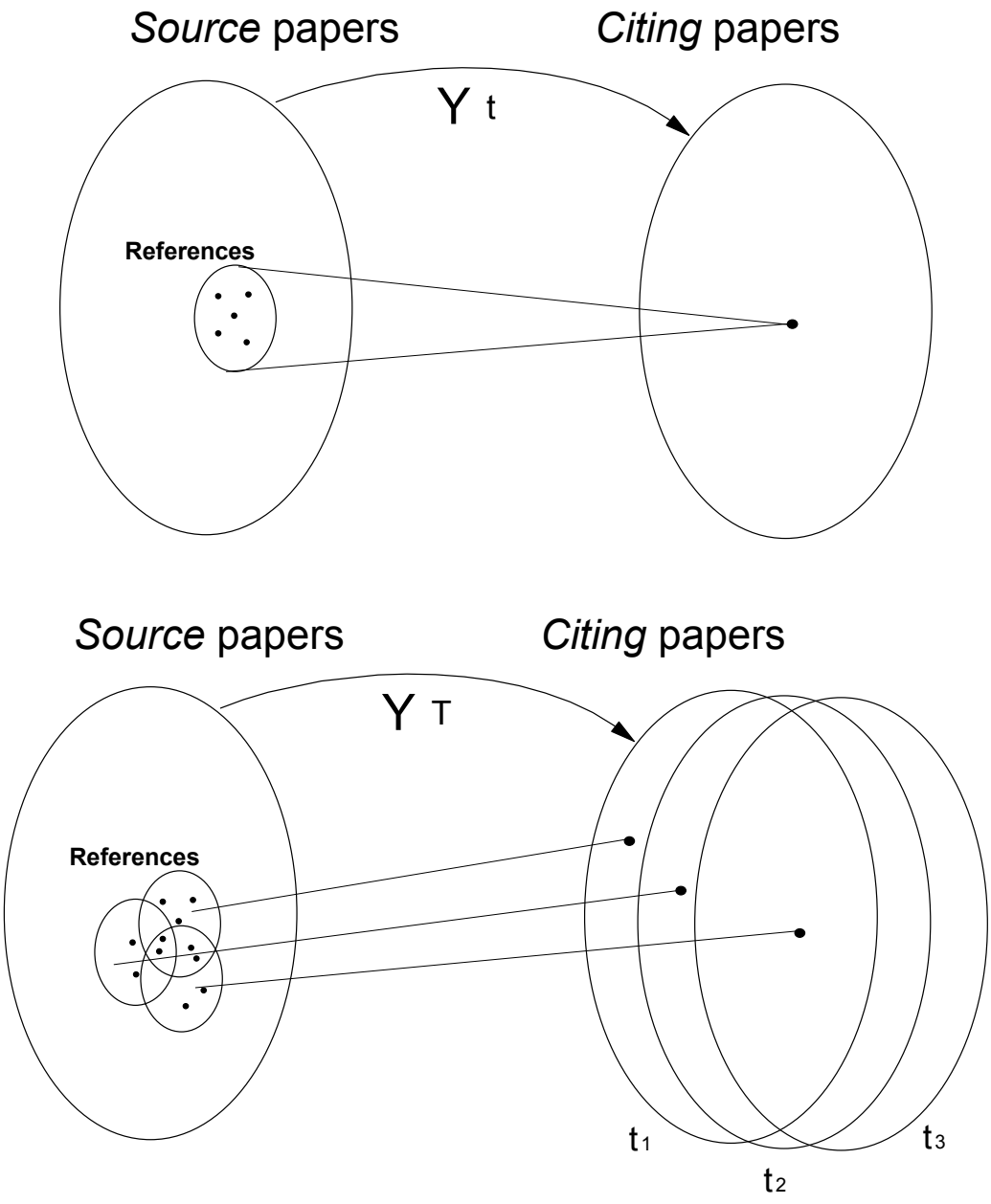



*Figure 2.8 The relationship between bibliometric elements and units in terms of their mathematical interpretation (citations links)*

Figures 2.7 and 2.8 visualise the formalism of mathematical interpretation of relationship between bibliometric elements. In the first case, the mapping $F_t$ (for each time $t$) is describes an *authorship*. $t$ is the publication date. The observation period $T = [s,\ \mathrm{t}]$ is called *publication period.* $F_T$ then forms the publication process of an author. The complete origin $F_t^{-1}$ of a given publications is the set of its *co-authors*, in particular, if we denote the set of authors by A and that of papers by B, then we have for each element $b \in B$: $F_t^{-1}(b) = \{a \in \mathrm{A}: F_t(b) = b\}$.

In the second case, the mapping $Y_t$ (for each time $t$) describes a *citation*. $t$ is the publication date of the citing paper. The observation period $T = [s,\ \mathrm{t}]$ is called *citation window.* $Y_T$ then forms the citation process of a paper. The complete origin $Y_t^{-1}$ of a given publications is the set of its *references*, in particular, if we denote the set of source papers by A and that of citing papers by B, then we have for each element $b \in B$: $Y_t^{-1}(b) = \{a \in \mathrm{A}: Y_t(b) = b\}$.

### *2.4.3 Deterministic models of productivity and citation processes*

#### *Publication activity as a measure of 'scientific productivity'*

As already mentioned in the context of the epidemic model according to Goffman, an open population has to be assumed since susceptible and infected persons have to be continuously replaced by persons entering the system. The model by *Schubert* and *Glänzel* introduced in 1983 describes similar publication processes. This model assumes three groups in the population, namely, 1. those who entering the system, 2. those who are staying in the system and 3. those who are leaving it. The system can be described as follows.
Consider an infinite array of units indexed in succession by the non-negative integers. The content of the $i$-th unit is denoted by $x_i$, the (finite) content of all units by $x$. Then the fraction $y_i = x_i/x$ ($i \geq 0$) expresses the share of elements contained by the $i$-th cell. The change of content is postulated to obey the following rules.

(i) Substance may enter the system from the external environment through 0-th unit at a rate $s$;

(ii) substance may be transferred unidirectionally from the $i$-th unit to the ($i$+1)-th one at a rate $f_i$ ($i \in \mathbf{N}_0$);

(iii) substance may leak out from the $i$-th unit into the external environment at a rate $g_i$ ($i \in \mathbf{N}_0$).

The next step towards a stochastic model is to interpret the above ratios $y_i$ as the (classical) probability with which an element is contained by the $i$-th unit. The stochastic process is then formed by the change of the content of the units, i.e., by the change of papers published by the authors who have entered the system. $X(t)$ denotes the (random) number of published papers, $P(X(t) = i) = y_i$ the probability that an author in the system has published exactly $i$ papers in the period $t$. Finally, the stochastic model itself is obtained if $X(t)$ is considered the *publication activity process* of an arbitrary author, and $P(X(t) = i) = y_i$ is the probability that he/she has published $i$ papers in the interval $(0, t)$. Figure 2.9 visualises the scheme of substance flow of this process.

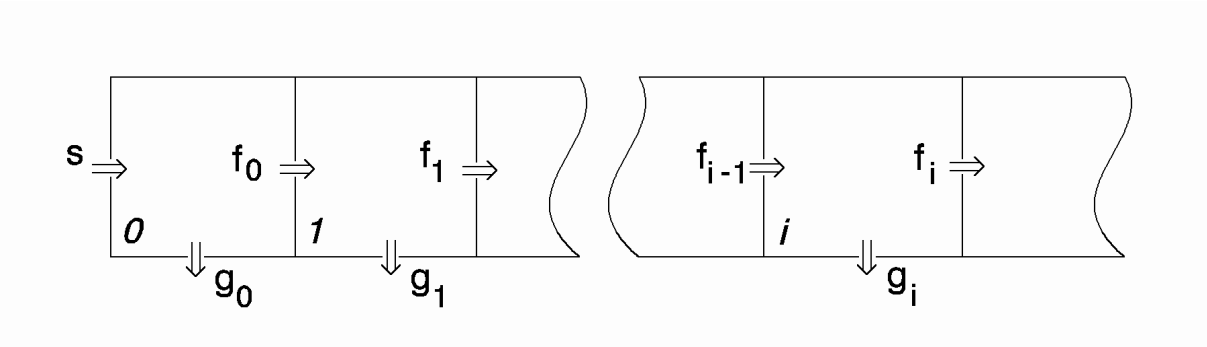



*Figure 2.9 Scheme of substance flow in the Waring process*

Now, using the above notations, we can give a mathematical formulation for the equations of change in the system.

$$\begin{aligned} x_0'(t) &= s(t) - f_0(t) - g_0(t) , \\ &\vdots \\ x_i'(t) &= f_{i-1}(t) - f_i(t) - g_i(t) , \quad (i > 0). \end{aligned} \tag{1}$$

Here and in the followings, the prime (') denotes time derivatives.
According to Schubert and Glänzel, the following particular forms of the above rate terms are used:

$$s = \sigma \cdot x , \tag{2}$$

$$f_i = (a + b \cdot i) \cdot x_i \; ; \; (i \geq 0) , \tag{3}$$

$$g_i = \gamma \cdot x_i \; ; \; (i \geq 0) , \tag{4}$$

where $\sigma$, $a$, $b$ and $\gamma$ are non-negative real values. Since $x'(t) = \Sigma x_i' = (s - \Sigma g_i) = (\sigma - \gamma) \cdot x$ (cf. Eq. (1)), the distribution of the substance over the units during time $t$ can be obtained as a solution of the following system of first order linear differential equations:

$$\begin{aligned} y_0'(t) &= \sigma - (a + \sigma) \cdot y_0 \\ &\vdots \\ y_i'(t) &= (a + b \cdot (i-1)) \cdot y_{i-1} - (a + b \cdot i + \sigma) \cdot y_i \; ; \; i > 0 , \end{aligned} \tag{5}$$

with the initial conditions

$$y_i(0) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise.} \end{cases}$$

For the entire population we can derive $x(t) = x(0)\cdot\exp((\sigma-\gamma)\cdot t)$, i.e., the system is asymptotically time-invariant (stationary) if $\sigma = \gamma$, otherwise, if $\sigma > \gamma$ or $\sigma < \gamma$ , it exponentially grows or decays, respectively. The general solution of the above system of differential equations is

$$y_i(t) = \sum_{j=0}^{i} b_{ij} e^{-(a+b\cdot i+\sigma)t} + \frac{\sigma(a+b)\ldots(a+b(i-1))}{(a+\sigma)(a+b+\sigma)\ldots(a+bi+\sigma)} \qquad (6)$$

where the coefficients $b_{ij}$ are determined by the initial conditions. Finally, if $\sigma = \gamma$ is assumed, the above process has a non-degenerated limit distribution, which is a *Waring* distribution with parameters $N = a/b$ and $\alpha = \sigma/b$ .

The following three special cases should be mentioned.

1. $a = 2$ $\Rightarrow$ *Price* distribution (*Glänzel & Schubert*, 1985)
2. $N = 1$ $\Rightarrow$ *Yule* distribution
3. $b = 0$ $\Rightarrow$ Geometric distribution

Derrek de Solla Price distinguished the following categories of authors: *newcomers, continuants, transients* and *terminators*. Within the framework of this model, $s$ represents the group of newcomers, $g_i$ terminators, $g_1$ transients and substance remaining in the system represents the group of continuants.
We just mention in passing that the two cases of the epidemic model according to Goffman are obtained if $\sigma = \gamma$, or $\sigma > \gamma$, respectively.

*Citation impact as a measure of reception of information*

The following model introduced by *Glänzel/Schubert/Schoepflin* (1994/95) uses a negative binomial process (a special case of a non-homogeneous birth-process) to describe the change of citation impact in time and for the ageing of scientific literature.
Consider now the same array of units as in the case of publication activity. In contrast to the above model we now assume that the system is completely isolated from external influences, i.e., no substance enters or leaves the system. Therefore only rule (ii) of the preceding paragraph remains valid. Hence $x(t) = x(0)$ follows immediately, where, of course, $x(0) > 0$ is assumed. The special assumption $x(0) = 1$ does not mean any restriction of generality. Now we reformulate Eqs (2)-(4):

$$\sigma = 0 \,, \qquad (2')$$

$$f_i = (a + bi)\cdot x_i \,;\; a(t)/b(t) = \text{const} \;(>0) \,, \qquad (3')$$

$$g_i = 0 \;;\; (i \geq 0) \,. \qquad (4')$$

The subsidiary condition to Eq. (3') says that the process is non-homogeneous, i.e., the substance flow may depend on the time elapsed. The proportionality coefficient in the subsidiary condition is denoted by $N$ , i.e., $a(t) = N\cdot b(t)$. Figure 2.10 shows the scheme of substance flow of this process.

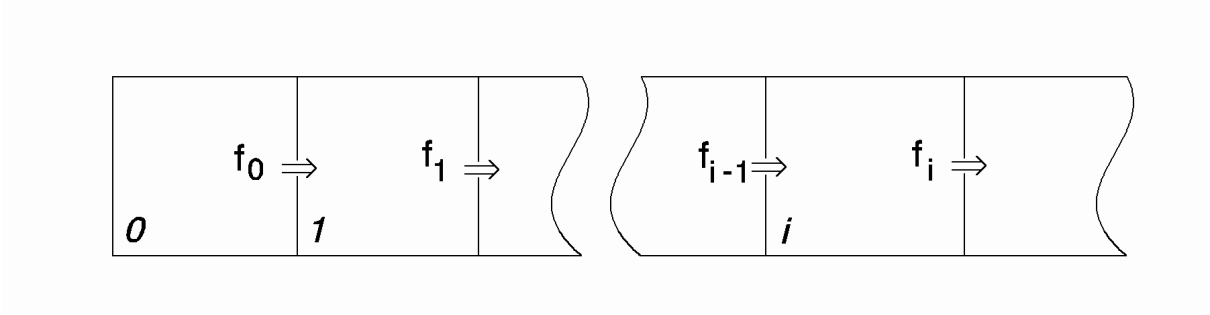



*Figure 2.10 Scheme of substance flow in the non-homogeneous birth process*

By analogy to the Waring model we have

$$\begin{aligned} y_0'(t) &= -N\cdot y_0\cdot b(t) \\ y_1'(t) &= (N\cdot y_0(t) - (N+1)y_1(t))\cdot b(t) \\ &\vdots \\ y_i'(t) &= ((N+i-1)\cdot y_{i-1}(t) - (N+i)y_i(t))\cdot b(t)\ , \end{aligned} \tag{5'}$$

with the same initial conditions as above.

Publication-activity dynamics is basically reflected by the distribution $P(X(t) = i)$ of the process. Special conditional probabilities, the so-called transition probabilities, however, permit a much deeper insight into scientific productivity processes. In our case, for example, the probability that an author publishes $j$ papers during $t$ years, provided he/she has published already $i$ papers during $s$ years, is called transition probability ($i \le j$, $s < t$). These probabilities reflect the influence of an initial period on the further publication activity. In particular, the probability that at time $t$ the substance is in the $k$-th unit, provided it was in the $i$-th one at time $s$, is denoted by $p_{ik}(s,t)$ ($k \ge i$). With the transition rules as above, we can write (see e.g., *Karlin* and *Taylor*, 1975)

$$\partial p_{ik}(s,t) / \partial t = \{(N+k-1)p_{i,k-1}(s,t) - (N+k)p_{ik}(s,t)\}\cdot b(t)\ ;\ \ k > i \tag{7}$$

The initial conditions are

$$p_{ik}(s,s) = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, let $X(t)$ denote the (random) index of that unit in which the substance is at time $t$. Then the following distribution can be obtained from the first system of differential equations by successive integration

$$P(X(t)=k) = y_k(t) = \binom{N+k-1}{k} e^{-r(t)N}(1-e^{-r(t)})^k, \tag{8}$$

where $r(t)=\int_0^t b(u)\mathrm{d}u$ , i.e., the distribution of substance over the units is negative binomial at any time. Analogously, the second system of differential equations results in

$$p_{ik}^{*}(s,t) = P(X(t) - X(s) = j | X(s) = i) = \binom{N+i+j-1}{j} \cdot e^{-(r(t)-r(s))(N+i)} (1 - e^{-(r(t)-r(s))})^{j} , \quad (9)$$

where $p_{ik}^{*}(s,t) = p_{ij}(s,t)$ with $i = k-j \geq 0$ . Eq. (9) can be reformulated verbally as follows. The substance flow during the time period $t$-$s$ has a negative binomial distribution with parameters $\exp(-r(t)+r(s))$ and $N+j$, where $j$ is the index of the unit which was reach by the substance at time $s$.
The mean value function of the process is defined as the regression function of $X(s, t)$ on $X = i$, namely,

$$M_i(s,t) = E(X(t)-X(s)|X(s)=i) . \quad (10)$$

This function will play an important role in the applications. Under the above conditions we have

$$M_i(s,t) = (N+i) (\exp(r(t)-r(s)) - 1) ; \quad i \geq 0, t \geq s \quad (11)$$

and

$$M(s,t) = E(X(t)-X(s)) = N\cdot(\exp(r(t))-\exp(r(s))) \quad (12)$$

Eq. (12) reflects the non-homogeneity of the process, i.e., for example, $M(s,s+h) \neq M(t,t+h)$ if $s \neq t$ ($h > 0$). Non-homogeneity is an important property of citation processes. The process has a non-degenerated or degenerated limit distribution according as $\lim r(t) < +\infty$ or $\lim r(t) = +\infty$ , respectively.

*A model of 'synchronous' and the 'diachronous' citation processes and the ageing of literature*

In citation studies, two basis approaches have been distinguished, namely, the *diachronous* and the *synchronous* model. The *diachronous* approach is concerned with the use of a given set of publications in successive years, whereas *synchronous* studies proceed from the present to the past. Consequently, we can derive two types of different citation processes, particularly, the 'synchronous' and the 'diachronous' process. In some recent publications, *Burrell* (2001, 2002) has used the terms *retrospective* and *prospective* citation studies, where the terms 'retrospective' and 'prospective' are practically synonyms for 'synchronous' and 'diachronous', respectively.

First *Wallace* (1986) has studied ageing of scientific literature. In particular, he analysed, the relationship between journal productivity and obsolescence, and assumed an exponential distribution. In his model, ageing analogously is related to the radioactive decay characterised by the "half-life" being the median of the distribution. Wallace's study is based on the *synchronous* approach, that is, he analysed the age of reference literature. From the pragmatic

point of view, on can say that synchronous analyses are easier to conduct since it does not require the observation of citations in a quite large citation window of ten, fifteen or even more years. Nevertheless, the synchronous approach cannot serve as a substitute for diachronous studies since the two approaches shed light on quite different aspects of citation processes, in general, and of ageing, in particular.
Although most *synchronous* (*retrospective*) ageing studies are based on the analysis of references in selected papers, *synchronous* studies can also be concerned with the analysis of citation received by publication sets. In both cases, the citation window is fixed and the publication period is variable. Thus, both the Citing and the Cited Journal Package in the annual up-dates of the Journal Citation Reports (JCR) have, for instance, to be considered *synchronous* citation approaches.
The terms *synchronous/diachronous* are nowadays also used in a more general context of citation analyses that are not directly related to ageing. *Ingwerson* et al. (2000, 2001) have introduced a distinction between *synchronous* and *diachronous* impact measures for scientific journals. According to their approach, the 'Garfield Impact Factor' produced by the Institute for Scientific Information is a synchronous Impact Factor as it the citation year is fixed and the two-year publication period lying in the "past". In the above-mentioned studies, *Ingwersen* et al. have given a methodological discussion why the 'diachronous' approach should be preferred to the 'synchronous' one. The fact that only diachronous impact measures can be calculated for non-serials is one of the advantages. The journal impact measures built since 1995 at LHAS in Hungary and at RASCI in Germany may serve as examples for such diachronous impact measures; here the publication year is fixed and citations are counted in a three-year observation period, namely, in the year of publication and the two subsequent years (for instance, *Glänzel*, 1996).

In the following, we will give a brief overview of ageing from the perspective of technical reliability visualising the different aspects that can be analysed by the two approaches. In the subsequent section a stochastic model will be given showing that one and the same model can be used to describe both *diachronous* (*prospective*) and *synchronous* (*retrospective*) perspectives of citation processes.

### *2.4.4 The stochastic approach to bibliometrics*

#### *Diachronous and synchronous ageing studies in the context of technical reliability*

Statistical functions provided by *synchronous* studies depend on too many factors to be uniquely interpreted in terms of ageing alone. In order to illustrate this effect we will refer to a simple model adopted from technical reliability processes (see, for example, *Watson* and *Wells*, 1961, *Gupta*, 1981). This may help to interpretation the two approaches. It should be stressed at once that the technical reliability model can not be applied to explain information processes, moreover, basic principles such as *lack of memory property* (technical reliability) and *cumulative advantage principle* (bibliometric processes) can be considered almost diametrical. The basic questions concerning *synchronous* and *diachronous* approach to obsolescence are, however, very similar in informetrics and technical reliability.
Within the framework of technical reliability analyses, function and lifetime of a system, a machine, a device or equipment is studied. The equipment may, for example, consist of machines in factories or of automobiles, but it might also be a simple device such as a rubber tyre or an ordinary electric switch. The usual definition of reliability of a system or a device is the probability that it will give satisfactory performance of its intended function for a

specified period under specific operating conditions. Using a set of systems or devices assumed to have identical parameters, reliability measures then give information about the following.

1. The probability that a given system or device will operate for a specific time
2. The number of failures that will occur over a specific period of time
3. The average time between failures
4. The expected lifetime of a system, machine or device provided it works already for a given period without failure

*Burrell* (2002) has already pointed to the fact that it might perhaps be unfortunate to think of citation as a "failure", however, this approach allows adopting model and terminology of technical reliability. Moreover, if ageing is measured by the number of received citations and the time elapsing between successive citations, the assumed "analogy" between failures in technical reliability and citations in the informetric ageing model does not at all seem to be absurd.

Lifetime is usually interpreted whether as time until failure or as time until death or destruction. In more general terms one could also consider 'lifetime' *the time during which a system is or can be used*; beyond this time it will not perform its intended function anymore. This definition can straightaway be applied to information science and bibliometrics as well. In particular, if citations are interpreted as *one important form of use of scientific information within the framework of documented science communication* (*Glänzel* and *Schoepflin*, 1999) then "technical reliability" of a scientific paper expresses the performance of its intended function, namely, that it is read and that it is has an impact on scientific research. The latter one can at least in part be measured by citations. Lifetime can consequently be interpreted as the period until it is not cited anymore. Analogously to a brand-new device that is not operating satisfactorily, a paper that is never cited can be considered not to give satisfactory performance of its intended function already when it was published.

The concept of technical reliability implies a *diachronous* approach. A device is produced, it is operating, and failures and finally death or destruction is observed. Despite the different background and the different interpretation of the diachronous and synchronous approach, it might, however, occur that the mathematical laws derived from the two approaches are very similar, so that the same formula could be used in both cases. In order to disprove this possible assumption, we will give a simple empirical evidence for the necessary distinction between the diachronous and the synchronous approach to ageing of scientific literature. The following example may just serve to visualise why diachronous bibliometric processes are not merely the "reflection" of the corresponding synchronous process. In order to show this, we have selected three journals representing three different subject fields, particularly, the journals *Cell* (Biosciences), *JACS* (Chemistry) and *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* (since 1986: *Probability Theory and Related Fields*) (Mathematics). Figure 2.11 presents the two lifetime curves simultaneously.

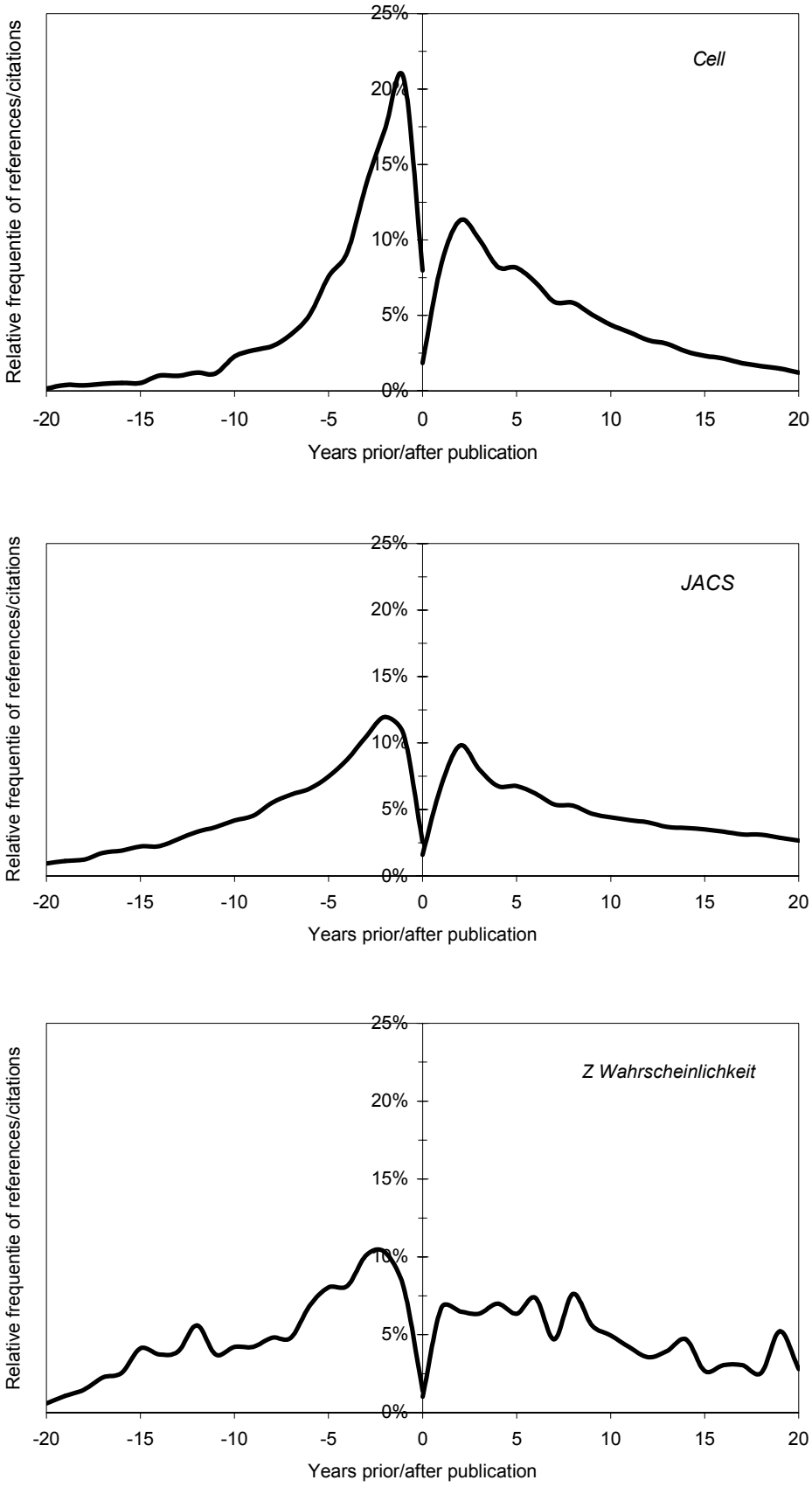



*Figure 2.11 Relative frequency of references and citations for three selected journals in 1980 (top: CELL, centre: JACS, bottom: Z Wahrscheinlichkeit)*

In what follows, a concise introduction into diachronous/synchronous citation processes will be given. It is based on a more general stochastic action-reaction models introduced by *Glänzel* (1983). In order to do without an excessive use of mathematical formalism, a verbal description of the rudiments is given. Nevertheless, the use of equations and formulae later on in the text will not completely be avoidable. The citation process will be defined as a diachronous process.

1. The basic idea of an action-reaction model is the introduction of stochastic processes through *timing-functions* defined on two discrete finite subsets in the original probability space. The two subsets will be denoted by **A** and **A'**, where these subsets need not be disjoint; they may even be identical. The elements of these sets are in the present paper assumed to be publications. The probability measure is defined as an elementary measure on the basis of a counting measure $\phi$.

2. There are two subsets **M** and **N** of **R** (the reel axis) or **Z** (the set of integers), depending on whether a continuous or discrete time model is used. Without loss of generality we assume that $\mathbf{M} \subseteq \mathbf{N}$. These sets represent the time-parameter. In particular, the mappings $\nu: \mathbf{A} \to \mathbf{M}$ and $\mu: \mathbf{A'} \to \mathbf{N}$ are called *timing-functions* indicating when an event has happened, that is, in the present case, when a paper is published. In the following, time is assumed to be discrete, that is, $\mathbf{M}, \mathbf{N} \subseteq \mathbf{Z}$. The elements of these sets can, instance, be the publication years of scientific papers.
3. The mapping $\tau: \mathbf{A'} \to \mathbf{A}$ describes the link between citing (**A'**) and cited papers (**A**). Using an a bit lax formulation, one can define its "inverse" mapping in the following manner. Let $\underline{\mathbf{A'}}_{\mathbf{N}}$ denote the set $\{\mathbf{A'}_{t_n,a} : t_n \in \mathbf{N}, a' \in \mathbf{A}\}$, where $\mathbf{A'}_{t_n,a} = \{a' \in \mathbf{A'}: \mu(a') = t_n \wedge \tau(a') = a\}$. In verbal terms, $\underline{\mathbf{A'}}_{\mathbf{N}}$ is the set of papers published in the period **N** and citing papers of the set **A**. Then the following mapping is uniquely defined: $\Lambda: \mathbf{N} \times \mathbf{A} \to \underline{\mathbf{A'}}_{\mathbf{N}}$, where $\Lambda(t_n , a) = \mathbf{A'}_{t_n,a}$ . The mapping $X^* = \phi \circ \Lambda$ then defines the *increments* of a stochastic process, namely the number of citations received by a paper $a$ at time $t_n$ . The process $X$ then is the number of citations received in the period $t \le t_n$ , that is, $X(t_n) = \Sigma_{t \le t_n} X^*(t)$ . $X: \mathbf{A} \times \mathbf{N} \to \mathbb{N}$ is an appropriate stochastic process that can be used as a model for *diachronous citation process*.

4. *Random selection.* It is clear that in the above paragraphs a stochastic process $X_s$ is defined for each $s \in \mathbf{M}$. The complete set of processes defined this way is $\underline{X} = \{X_s\}_{s \in \mathbf{M}}$ . Then the measurable mapping $\nu: \mathbf{A} \to \mathbf{M}$ defines a *random selection* of processes from the set $\underline{X}$. Without loss of generality we can extend the definitions of the processes by putting $\mathbf{N} = \mathbf{M}$ and defining $X_s(t) \equiv 0$ for all $t < s$. In the discrete case, a process $X_s(t_n)$ can then be selected with probability $p_s = P(\nu = s)$ for all $s \in \mathbf{M} = \mathbf{N}$. An alternative presentation of probability measures based on these processes is possible through the use of conditional distributions, where the condition is given by the selection of publication period, that is, for instance, $E\, X_s(t) = E(X(t) \mid \{\nu = s\})$.

5. The appropriate *synchronous process* can now readily be derived from the *diachronous* one by fixing $t$ and through random selection from the corresponding set

of increments $\underline{X}^* = \{X_s^*\}_{s \in \mathbf{M}}$. Random selection from the set $\underline{X}$, or from $\underline{X}^*$ with variable $t$, respectively, will result in *hybrid* diachronous-synchronous processes.

We give three examples for functions that can be used to measure ageing properties of the process. First, we can assume that if inf sup $\mathbf{N} = \infty$ and the process is convergent, that is, $X_s(t_n)$ converges to a non-degenerate random variable $X_s(\infty)$ with probability 1 for each $s \in \mathbf{M}$. We can further more assume that the sequence is uniformly integrable. Under these conditions, which are quite natural for citation processes, we can define the following measures.

The *mean-value function* is the simplest of these measures. It is defined as the mean value of $\mathrm{E}\, X_s(t_n)$ for each $s \in \mathbf{M}$ at any given time $t_n$. Its basic properties are obvious from the definition and the above assumptions: $\mathrm{E}\, X_s(t_n)$ is a non-negative, increasing function and $\mathrm{E}\, X_s(t_n) \to \mathrm{E}\, X_s(\infty) < \infty$.

Hence we can define the second function, the *life-time function* of the process (see *Glänzel*, 1983). In particular, it takes the following form:

$$\mathrm{P}(\mu \le t_n) = \frac{\mathrm{E}\, X_s(t_n)}{\mathrm{E}\, X_s(\infty)} \ .$$

This life-time function can also be called life-time distribution, since $0 \le \mathrm{P}(\mu \le t_n) \le 1$ and $\lim_{tn \to \infty} \mathrm{P}(\mu \le t_n) = 1$. This distribution has already been used, for instance, in the studies by *Glänzel* and *Schoepflin*, 1994 and *Burrell*, 2002.

One possibility to define an *obsolescence function* of the process $X_s$ ($s \in \mathbf{M}$) as the third function is given in the following (see *Glänzel* and *Schoepflin*, 1994). We say an element in the model is *obsolete* at time $t_n$ if it will not be cited at any time $t' > t_n$. The probability $H_s(t_n)$ of becoming obsolete is called *obsolescence function* of the process and is defined as

$$H_s(t_n) = \mathrm{P}(X_s(\infty) - X_s(t_n) = 0) \ .$$

$H_s$ is an increasing function on $\mathbf{N}$. Although $0 \le H_s(t_n) \le 1$ and obviously $H_s(\infty) = 1$, $H$ is in most cases not a distribution function since $H_s(s) = \mathrm{P}(X_s(\infty) = 0)$ and $\mathrm{P}(X_s(\infty) = 0)$ might be positive.

*Citation Distributions and Statistical Reliability of Comparisons*

First, we give an example for citation distributions of scientific journals without proposing any particular model. Citation distributions are discrete and often very skewed. The following figure (Fig. 2.12) presents the distribution of citation over papers published in the journals *Angewandte Chemie* (Intern. Edition) and *JACS* in 1995/96. Citations have been counted in the period 1995-1997 for papers published in 1995 and in 1996-1998 for papers published in 1996.

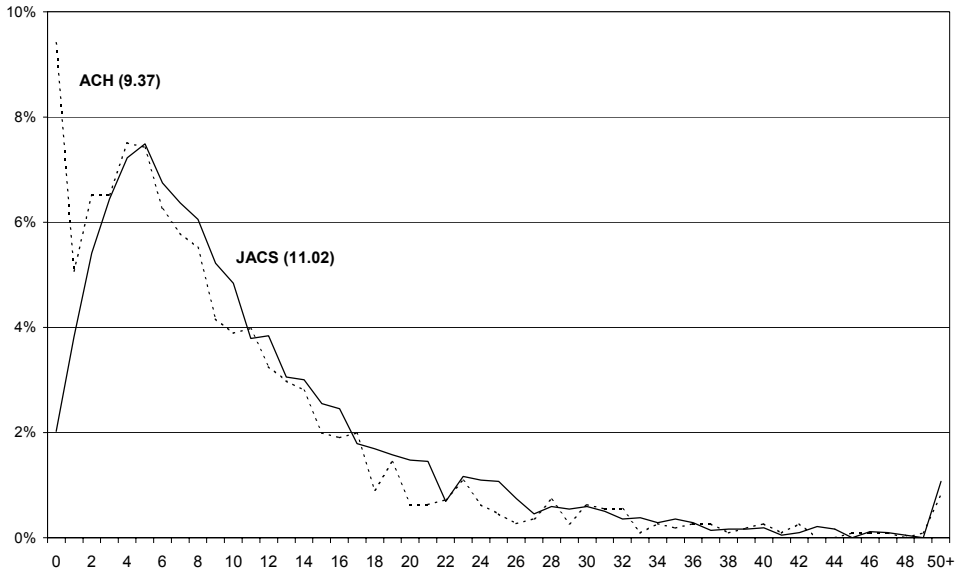



*Figure 2.12 Distribution of citation over papers published in the journals Angewandte Chemie (Intern. Edition) and JACS in 1995/96*

Figure 2.13 shows the empirical distribution of three samples and their fit on the basis of the above-mentioned model by *Glänzel*, *Schubert* and *Schoepflin*. Again, citations have been counted in the period 1995-1997 for papers published in 1995 and in 1996-1998 for papers published in 1996. The tree samples represent three different types of citation distributions. The first one (*The Lancet*) is extremely skewed and highly polarised. In a later section, we will see that that is in part due to the high share of *Letters to the editor* published in this journal. The shape of the distribution of the second sample (the field of *neuroscience*) is less skewed and less polarised. The last example, that of the journal *Nature* reflects the most advantageous situation.

It is a misbelief that the use of mean values is not justified in case of discrete, skewed distributions. Sample means have under certain conditions a $N(m, \sigma)$ distribution where $m$ is the expectation of the underlying (discrete) distribution and $\sigma$ depends only on the standard deviation of the distribution and the sample size. In this context, the Impact Factor and related measures based on other publication periods and citation windows can be interpreted as sample means. Then the *Mean Citation Rate* $\bar{x}$ of a journal is an unbiased estimate of the expectation of the underlying random variable $X$. That is, we have $E(\bar{x}) = E(X)$ and $D^2(\bar{x}) = D^2(X)/n$, where $n$ is the number of source items of the journal $J$ in the publication period. $D(\bar{x})$ is the standard error of $\bar{x}$. It can be used as a basis for comparisons with the impact of other journals and allows to decide whether the deviation of the impact of a journal from that of another one is significant or not (see *Schubert* and *Glänzel*, 1983). The following example might illustrate this.

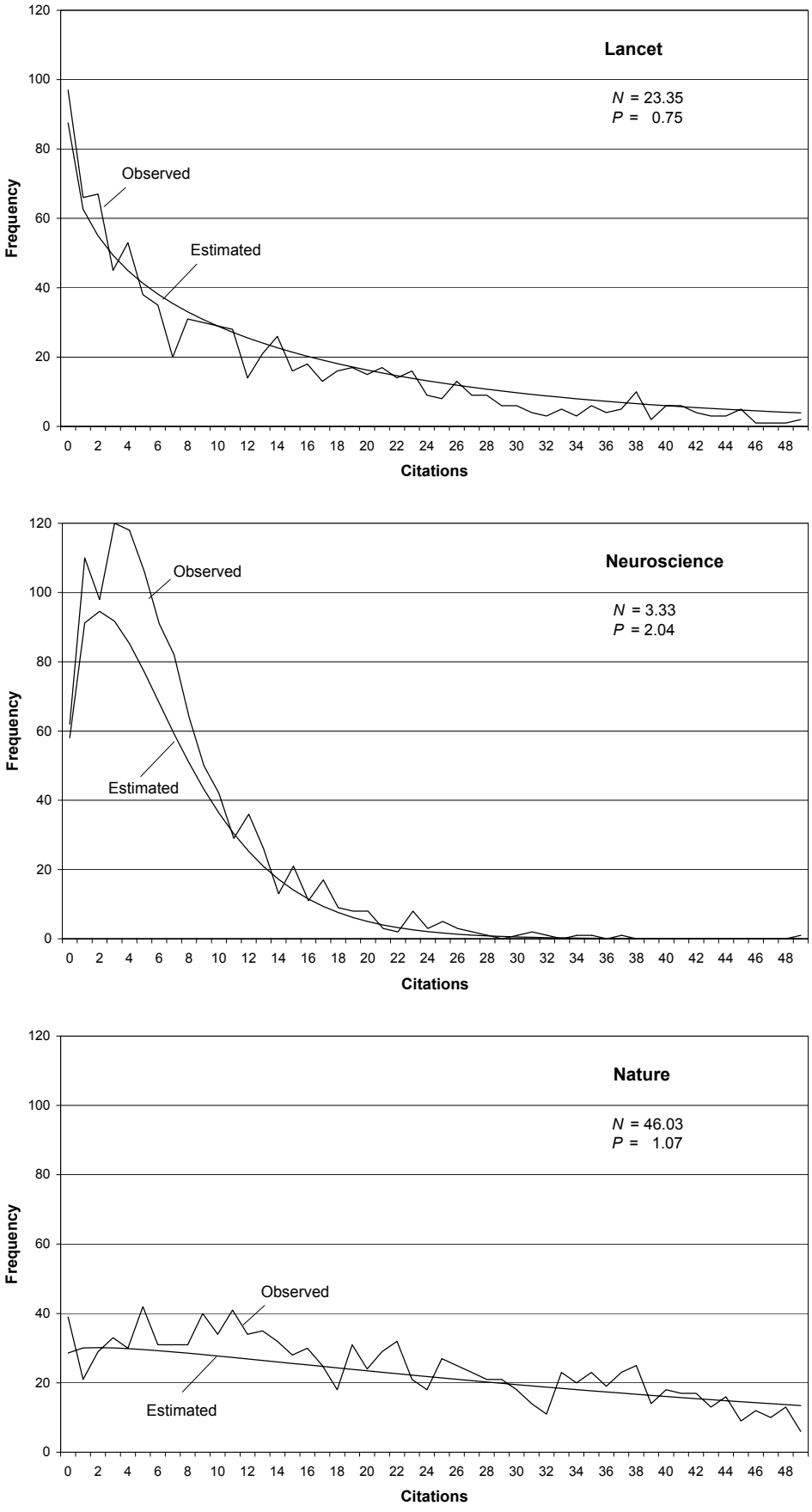



*Figure 2.13 Distribution of citation over papers and their fits on the basis of a negative-binomial distribution*

*Example:* Compare the mean citation rate of Austrian and Hungarian mathematical papers published in 1978/79 (citations counted in 1980).

| | | |
|---|---|---|
| Austria: | $x_A = 0.248$ | $D(x_A) = 0.073$ |
| Hungary: | $x_H = 0.226$ | $D(x_H) = 0.045$ |

We have

$$w = \frac{x_A - x_H}{\sqrt{D^2(x_A) + D^2(x_H)}} = \frac{0.248 - 0.226}{\sqrt{0.073^2 + 0.045^2}} = 0.257 .$$

Since $|w| < 1.96 = w_p$ ($p = 0.95$) the deviation is not significant at a confidence level of 0.95.

In this context, we have to mention that *Haitun* (1989) rejects the application of traditional mathematical statistics to bibliometrics referring to *non-Gaussian nature* of bibliometric data. However, his approach remained controversial.

Using stochastic models, further important questions can be answered, for instance, in how far the *succession of citations* received can be described by mathematical laws (e.g., *Glänzel*, 1992a, *Burrell*, 2001). Especially, the *first-citation distribution* is of particular interest. This has been studied by *Schubert* and *Glänzel* (1986), *Glänzel* (1991), *Rousseau* (1994), *Egghe* (2000) and *Burrell* (2001). *Schubert* and *Glänzel* (1986) have used first-citation to measure the speed of reception of scientific results.

# 3. INDICATORS OF PUBLICATION ACTIVITY

## *3.1 Counting schemes and main levels of aggregation*

Publication activity is expressed by the number of papers published by a selected unit in a given time.

Counting schemes are the method according to which publications are to be assigned to the contributing units. In particular, three "counting schemes" are used for publications.

1. The **fractional counting scheme**, that is, if *n* units (authors, institutions, countries, etc.) have contributed to the paper in question, each contributing unit takes the value 1/*n* for this paper (for instance, applied by *CHI Research Inc.*, Haddon Heights, NJ, USA)
2. The **first address count**, i.e., a paper is assigned to one unit only, on the basis of the first address in the address list of a paper as included in the database (for instance, used by *Information Science and Scientometrics Research Unit* (ISSRU), Budapest, Hungary in the 80s).
3. The **full or integer counting** scheme assigns a co-publication fully to each contributing unit (applied by the *Centre for Science and Technology Studies* (CWTS), Leiden, Netherlands, in the 90s also by ISSRU and the *Research Association for Science Communication and Information e. V.* (RASCI), Germany).

Due to intensifying scientific collaboration, the use of the first or correspondence address proved not an appropriate scheme for the 90s. Although fractional counts are partially additive, i.e., data from a lower level of aggregation can be summed up a higher one or the total, this scheme proved highly problematic if the fractionation criterion (institutional, regional, national fractionation) is not documented, and data are taken out of their context. This can be illustrated by the following example presented in figure 3.1. This paper has nine co-authors working at three different institutions that are located in two different countries. Applying fractional counting to the example in Figure 3.1, each co-author contributes with a "share" of 0.111, each institution involved with 0.333 and each country with 0.500. Since the authors are not employees of the same institutions and the corporate addresses are in two different countries, fractional counts cannot be summed up among the different levels of aggregation. Figure 3.2 visualises what happens if subsets do not form partitions and different levels of aggregations (units) are represented by systems of overlapping subsets. The same formalism representing author-publication links is used as in Figure 2.7.

```
SCI CDE with Abstracts  (Jan 93 - Jul 93)       (D4.0)
Authors:      Prassides-K Kroto-HW Taylor-R Walton-DRM David-WIF Tomkinson-J Haddon-RC
              Rosseinsky-MJ Murphy-DW
Title:        Fullerenes and Fullerides in the Solid-State - Neutron-Scattering Studies
Full source:  CARBON 1992, Vol 30, Iss 8, pp 1277-1286
Addresses:    UNIV-SUSSEX, SCH CHEM & MOLEC SCI, BRIGHTON BN1-9QJ, E-SUSSEX, ENGLAND
              RUTHERFORD-APPLETON-LAB, DIDCOT OX11-0QX, OXON, ENGLAND
              AT&T-BELL-LABS, MURRAY-HILL, NJ 07974, USA
```

*Figure 3.1 Bibliographic data of the example for different fractional counting according to different levels of aggregation*

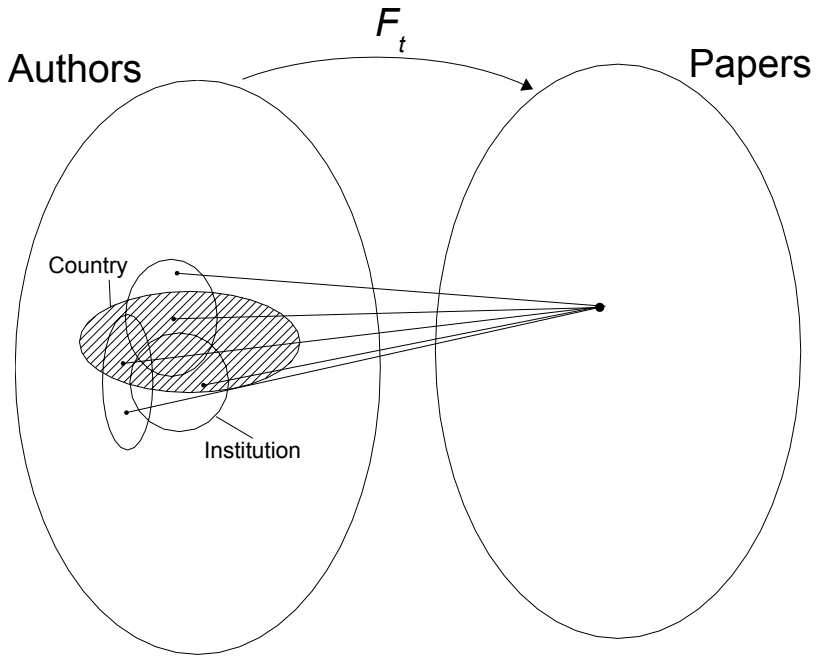



*Figure 3.2 Example for different levels of aggregations represented by overlapping systems of subsets*

As a consequence of this effect, comprehensive studies including analysis of collaboration patterns, comparisons of relative publication activity and relative citation impact require full-address counts. Moreover, each level of aggregation has its own bibliometric standards. Users should therefore take into consideration that data sets designed for studies at different levels of aggregation might be incompatible (see *Glänzel*, 1996).

*Levels of aggregation in bibliometric research*

From the viewpoint of bibliometric methodology, the distinction between three levels of aggregations is important. Each level of aggregation requires its own methodological and technological approach.

*Micro level*: Publication output of individuals and research groups

*Meso level*: Publication output of institutions; studies of scientific journals

*Macro level*: Publication output of regions and countries; supra-national aggregations

There are various reasons for this necessary distinction, among others, the mathematical-statistical background, the precision of retrieval and cleaning-up of data, different counting schemes, different meaning of bibliometric conceptions (e.g., self-citations), non-additivity of bibliometric data (because of multiple assignment).

*3.2 Problems of subject assignment*

Subject assignment is necessary to describe publication activity in given subject areas or sub-fields. Publications can be assigned to subjects through

a) subject codes or subject headings,
b) key-words or
c) journals classification.

Subject assignment is largely determined by the bibliographic database used for bibliometric application. From the viewpoint of the *given practical purpose*, two different basic schemes are used: Hierarchic and fine-structured classification systems used in information retrieval and more "robust" schemes emphasising science organisation aspects and science policy needs. Specialised databases such as INSPEC (PACS codes) or Mathematical Reviews (MSC codes) allow retrieval at very low levels of classifications, for very specialised topics. Figure 3.3 illustrates the hierarchical structure of the MSC-code system of the database "Mathematical Reviews" defining the topic "Characteristic functions; other transforms" as subtopic of "Distribution theory" which is in turn a sub-discipline of the subject "Probability theory and stochastic processes".

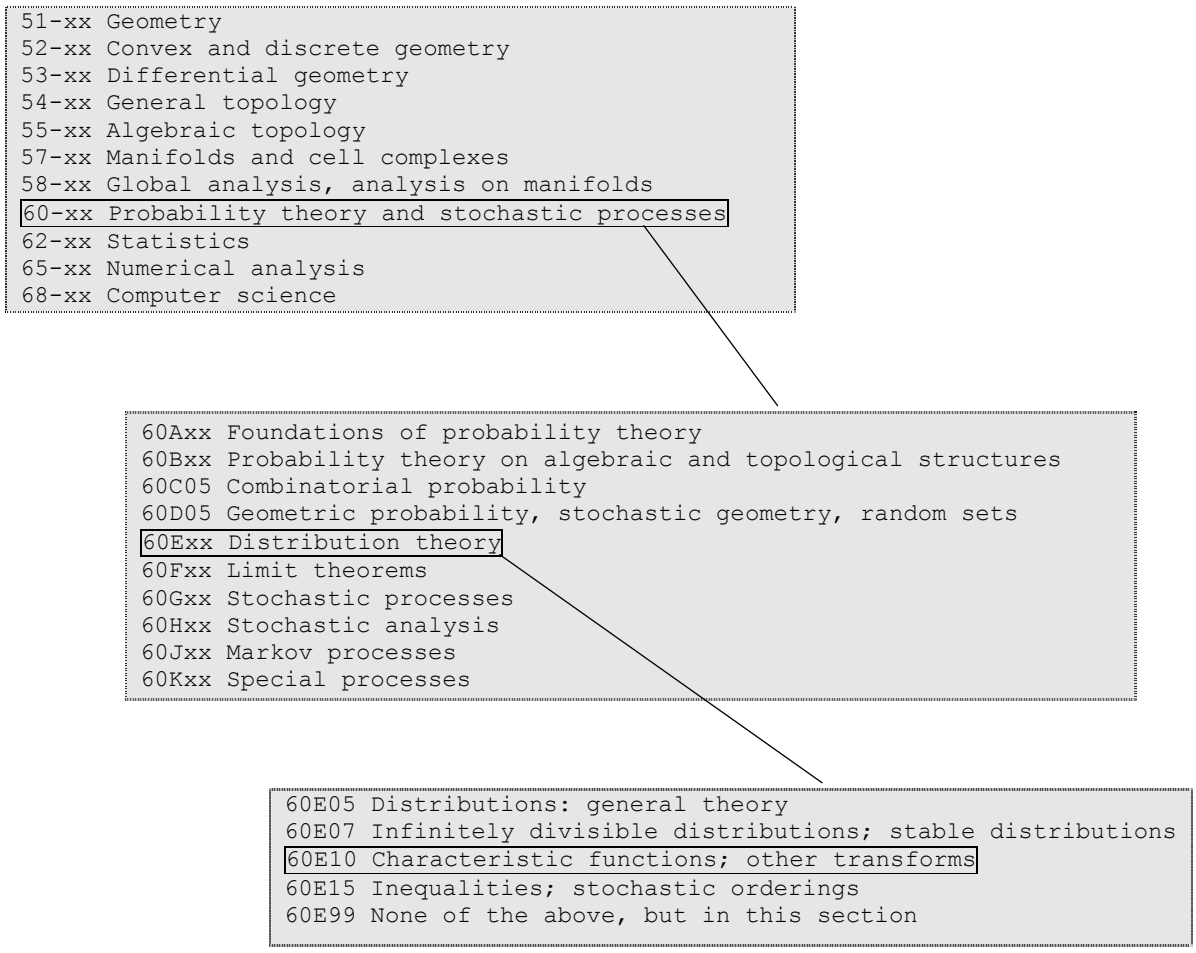



*Figure 3.3 Example for the hierarchical structure of the MSC-code system (subject 60: Probability theory)*

"Reliability" and "precision" of this scheme are excellent. Papers are assigned individually, on the basis of the authors' classification and corrections by reviewers. Nevertheless, subject classification (especially at lower hierarchical levels) is not unique as multiple assignments are allowed (MSC Primary/Secondary). Using the above example, an article might, for instance, be concerned with characteristic functions or other transforms of stable distributions (60E10 and 60E07).

As already mentioned, ISI databases do unfortunately, not provide a direct subject assignment for indexed papers. The multidisciplinarity of the database can be considered the main reason for lacking *paper-based* subject classification. The annual *Science Citations Index Guides* and the *ISI Journal Citation Reports* (*JCR*), however, contain regularly updated lists of journals assigned to one or more subfields (*ISI Subject Categories*) each. For lack of an appropriate subject-heading system, more or less modified versions of this *Subject Category* scheme are often used in bibliometric studies, too, namely as an indirect subject assignment of individual papers through those journals in which they have been published. Such assignment system based on journal classification has been developed among others by *Narin* and *Pinski* (see, for instance, *Narin*, 1976). Taking into consideration that journals are often not devoted to a single topic, the delimitation of subject areas based on journal assignment is necessarily less precise than that based on subject headings of individual publications as, for instance, in the MSC-code system. The specialisation of journals in Condensed Matter Physics (CMP) may just serve as an example. The extent of specialisation here ranges from *Act Chryst C* which is devoted to one single subject to *Phys Rev L* publishing papers in all the 17 CMP subdivisions according to the *Physics Abstracts* database (see *Todorov* and *Glänzel*, 1990). Since the above-mentioned approach by *Narin* and *Pinski*, other institutions have attempted to create their own hierarchical journal-based classification systems. In Europe, main three schemes developed at ISSRU (Budapest), ISI (Karlsruhe) and CWTS (Leiden) were in use. All these schemes are based on ISI journal assignment. Besides "general" journals that can at least be assigned to one of the major science fields like physics or chemistry, there are the big multidisciplinary journals such as *Nature, Science* and *P NAS US* publishing papers from almost all disciplines. A journal assignment would here obviously fail. The same applies to many selectively covered journals in the SSCI and AHCI. Selectively covered journals are often represented by one or two papers that are indexed because all other papers published in the same issue are not relevant for the corresponding database. In order to overcome this problem, several methods have been developed to assign papers published in such journals individually. Among those, the following two approaches deserve to be mentioned. 1. using cognitive words, for instance, from the address field (*de Bruin* and *Moed*, 1993) and 2. analysing the reference literature of those papers (for instance, *Glänzel* et al., 1999).
Figure 3.3 may serve as an example for a classification scheme developed for bibliometrics purposes on the basis of journal assignment and applying an individual scheme based on reference literature to papers published in multidisciplinary journals. The scheme can be characterised as a two-level hierarchical system for subject fields and subfields and keeping the subject categories still as the third and lowest level. The example presented in Figure 3.3 shows the subject category "gerontology" at a category of the subfield "age & gender related medicine" within the field "non-internal medicine". This type of classification schemes proved to be robust enough for policy-relevant bibliometric applications.

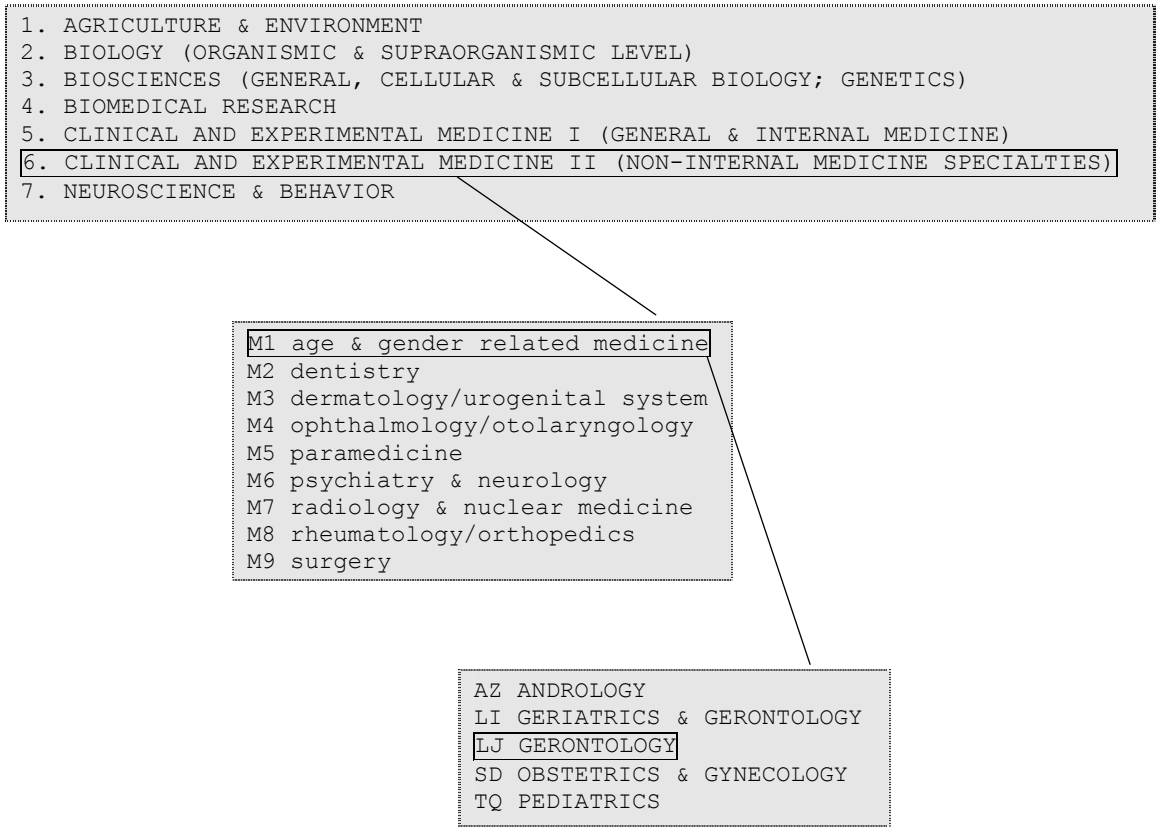



*Figure 3.3 Example for the hierarchical structure of a scheme based on ISI categories (Subject field: non-internal medicine)*

As already mentioned earlier, assignment to units (for instance, on the basis of addresses) is often not unique. The same applies to subject classification. A problem arises if data have to be aggregated or disaggregated from one level of aggregation to another one. Whenever publications cannot be uniquely assigned to one single category aggregation or disaggregation is problematic or may lead to invalid results. We have to stress again that bibliometric indicators are *not additive* over (sub-)fields or most other units of aggregation. Alone journals always form disjoint sets, so that measures based on mere journal data are usually additive.

### *3. 3 Statistics on scientific productivity: Frequency distributions vs. rank statistics*

*Alfred J. Lotka* was the first scientist who tried to find regularity in the publication activity of authors of scientific publications. In 1926, he published his pioneering study on the frequency distribution of scientific productivity determined from a decennial index (1907-1916) of *Chemical Abstracts. Lotka* concluded "the number (of authors) making $n$ contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that makes a single contribution, is about 60 per cent." Minimum productivity in this model in 1, the number of authors producing one single paper is denoted by $n_1$, that of authors with maximum productivity is $n_{max}$ where $n_{max} > 1$ or $n_{max} = 1$, according as there is a tie at the first rank or not.

Lotka's Law is closely linked with another important law formulated by de Solla Price (1963), namely, that

> "*half of the scientific papers are contributed by the top square root of the total number of scientific authors.*"

This law is called *square root law*. *Allison* et al. (1976) have shown that the square root law follows from Lotka's Law only if an additional and not unique assumption is made about the number of contributions ($n_{max}$) by the most active author.

From Lotka's Law, the following distribution has been derived:

$$P(X = k) = (6/\pi^2)\cdot k^{-2} \text{ , } k = 1, 2, 3\dots \text{ ,}$$

that is, no maximum productivity is assumed, or, in other words, the maximum might be infinite. Unfortunately, Lotka's Law does not take the possibility of temporary inactivity $\{X = 0\}$ into account. Nevertheless, this result can be considered as a rule of thumb even today, 75 years after its publication.
In 1985, *Glänzel* and *Schubert* have proved that a special case of the *Waring* distribution satisfies the *square root law* according to Price without any further condition. Moreover, the suggested distribution also allows temporary inactivity as the probability of event $\{X = 0\}$ is defined. *Glänzel* and *Schubert* have called their distribution *Price distribution.* Being a special case of the *Waring* distribution, it can also be derived from the Waring process as described in Section 2. The probabilities of the Price distributions are as follows

$$P(X = k) = N\cdot\{1/(N+k) - 1/(N+k+1)\} \text{ , } k = 0, 1, 2, 3\dots \text{ ,}$$

Or, if we exclude the event $\{X = 0\}$

$$P(X = k \mid X > 0) = (N+1)\cdot\{1/(N+k) - 1/(N+k+1)\} \text{ , } k = 1, 2, 3\dots \text{ .}$$

| k | Observed frequency | Calculated frequency | |
|---|---|---|---|
| | | *Lotka* | *Price** |
| 1 | 3991 | 3991.0 | 3991.0 |
| 2 | 1059 | 997.8 | 1063.6 |
| 3 | 493 | 443.4 | 492.8 |
| 4 | 287 | 249.4 | 284.3 |
| 5 | 184 | 159.6 | 185.0 |
| 6 | 131 | 110.9 | 130.0 |
| 7 | 85 | 81.4 | 96.3 |
| 8 | 64 | 62.4 | 74.3 |
| 9 | 65 | 49.3 | 59.0 |
| 10 | 41 | 39.9 | 48.0 |
| >10 | 491 | 705.9 | 466.8 |

*Figure 3.4 Frequency distribution of scientific productivity according to Lotka (k = number of papers, Price distribution with N = 0.727)*

Figure 3.4 presents the original Lotka data as well as the fit on the basis of the Lotka and the Price distribution according to *Glänzel* and *Schubert*, 1985).

One year later, in 1986, *Egghe* and *Rousseau* have presented another solution for Price's square root law.

Another approach to distributional presentation and analysis of productivity data are *rank frequencies*. Figure 3.5 presents the rank frequency table of the distribution of words in Joyce's Ulysses according to *Zipf* (1949). According to Zipf's Law, the product of rank and frequency *rf* should be constant *C* (in this example $C \sim 24{,}500$). *Mandelbrot* (1963) has shown that a law of the form $f = A\{1 + Br\}^{-\alpha}$ is more appropriate. Nowadays, the latter form is used whenever Zipf's Law is referred to.

| *r* | *f* | *r·f* |
|---:|---:|---:|
| 10 | 2,653 | 26,530 |
| 20 | 1,311 | 26,220 |
| 30 | 926 | 27,780 |
| 40 | 717 | 28,680 |
| 50 | 556 | 27,800 |
| 100 | 265 | 26,500 |
| 200 | 133 | 26,600 |
| 300 | 84 | 25,200 |
| 400 | 62 | 24,800 |
| 500 | 50 | 25,000 |
| 1,000 | 26 | 26,000 |
| 2,000 | 12 | 24,000 |
| 3,000 | 8 | 24,000 |
| 4,000 | 6 | 24,000 |
| 5,000 | 5 | 25,000 |
| 10,000 | 2 | 20,000 |
| 20,000 | 1 | 20,000 |
| 29,899 | 1 | 29,899 |

*Figure 3.5 The distribution of words in Joyce's Ulysses illustrating Zipf's law (r = rank, f = absolute frequency)*

*Bradford* (1934) discovered his regularity when studying the extent to which literature in a single discipline is scattered over a range of journals. He found that "if scientific journals are arranged in order of decreasing productivity on a given subject, they may be divided into a nucleus of journals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus when the numbers of periodicals in the nucleus and the succeeding zones will be as 1: $b$ : $b^2$ …". Let $N_n$ denote the number of journals in the $n$-th zone and $N_0$ the number of journals in the core. If the core and each zone contained the same number of articles then $N_n = k^n N_0$, where $k$ is the so-called *Bradford coefficient* (denoted by $b$ in the above formulation by Bradford). *Leimkuhler* (1967) has shown that this law can be reformulated as a particular *rank frequency law*. *Bookstein* (1990) has shown that Leimkuhler's form is stronger than Bradford' since Bradford only claimed that a core can be found obeying his regularity. If we proceed from Leimkuhler's law, and allow

the notion of fractional journals and articles then the Bradford regularity will hold. In the Leimkuhler version, a distinguished core is, however, not required.

We will just sketch the proof of the equivalence of the two regularities. The basic idea is the total number of journals $N$ can be obtained by summing up over zones resulting in the following equation.

$$N = \Sigma\, N_i = \Sigma_{0 \le i < n}\, k^i N_0 = N_0\,(1 + k + k^2 + k^3 + \ldots + k^{n-1}) = N_0(k^n - 1)/(1 - k),$$

Where we assume that we have $n$ zones. Let $Y_0$ denote the number of papers belonging to each zone (according to Bradford, this number is constant) and $Y$ the total number of papers. Consequently, we have $Y = nY_0$. Hence we have

$$Y/Y_0 = n = \ln\,\{1 + (k - 1)N/N_0\}/\ln\, k\,.$$

Putting $b := k–1$ yields the form given by Leinkuhler. The reverse of this derivation holds only if $Y/Y_0$ is an integer.

| A | B | C | D |
|---|---|---|---|
| 1 | 93 | 1 | 93 |
| 1 | 86 | 2 | 179 |
| 1 | 56 | 3 | 235 |
| 1 | 48 | 4 | 283 |
| 1 | 46 | 5 | 329 |
| 1 | 35 | 6 | 364 |
| 1 | 28 | 7 | 392 |
| 1 | 20 | 8 | 412 |
| 1 | 17 | 9 | 429 |
| 4 | 16 | 13 | 493 |
| 1 | 15 | 14 | 508 |
| 5 | 14 | 19 | 578 |
| 1 | 12 | 20 | 590 |
| 2 | 11 | 22 | 612 |
| 5 | 10 | 27 | 662 |
| 3 | 9 | 30 | 689 |
| 8 | 8 | 38 | 753 |
| 7 | 7 | 45 | 802 |
| 11 | 6 | 56 | 868 |
| 12 | 5 | 68 | 928 |
| 17 | 4 | 85 | 996 |
| 23 | 3 | 108 | 1,065 |
| 49 | 2 | 157 | 1,163 |
| 169 | 1 | 326 | 1,332 |

*Figure 3.8 Bradford's data on applied geophysics*
*(A = number of journals producing the corresponding articles in column B, B = number of relevant papers found in each journal, C = journal rank in descending order of productivity and D = cumulative number of papers)*

If we are not summing up $N_n = k^n N_0$ over all zones but just over the first $r$ ($r \le n$) zones we obtain analogously to the above derivation, the following important rank "distribution":

$$R(r) = Y_0 \cdot \ln \{1 + (k-1)r/N_0\}/\ln k \,,$$

where $R(r)$ is the cumulative number of papers in the first $r$ journals. For $r = n$, we have, of course, $R(r) = Y$. Putting $A := Y_0/\ln k$ and $B := (k-1)/N_0$, we obtain the following known form of Leimkuhler's law.

$$R(r) = A \cdot \ln (1 + B \cdot r) \,, \; r = 1, 2, 3 \ldots \,.$$

Figure 3.6 presents the original data by Bradford's on applied geophysics.

*Rank statistics and ordered samples*

An elegant approach to rank statistics is that using Gumpel's characteristic extreme values (*Gumbel*, 1958). Consider a given sample $\{X_i\}_{i=1,\ldots,n}$. Assume that the sample is ranked in descending order, that is,

$$X_1^* \ge X_2^* \ge \ldots \ge X_n^* .$$

In the "quantile" approach, the statistic $X_k^*$ ($1 \le k \le n$) is considered the $(1 - k/n)$-quantile. Gumbel has defined the characteristic $k$-th extreme values as

$$u_k = G^{-1}(k/n) = \sup\{x: G(x) > k/n\}; \; k = 1, 2, \ldots, n,$$

where $n$ is the sample size, $G = 1–F$ and $F$ is the common cumulative distribution function of the random variables $X_k$. *Glänzel* and *Schubert* (1988) have shown that there is a sequence of real values $\lambda_k \in (k–1, k)$ independent of the distribution of the sample, so that

$$P(X_k^* < u_k^*) = P(X_k^* < G^{-1}(\lambda_k/n)) = 0.5 \,,$$

that is $2\lambda_k$ can be considered the median of a $\chi^2$-distribution with $2k$ degrees of freedom. Then the *modified* Gumbel's characteristic extreme values $u_k^*$ are the median of the corresponding ordered statistics $X_k^*$. $\lambda_k$ can be approximated by $\lambda_k \sim (k - 0.3)$ and we have $\lim_{k\to\infty} (k - \lambda_k) = \ln 2$.
Two types of statistical tests can be derived, particularly,

1. a multi-sample test, for instants, if the extreme values (e.g. the maxima) of sub-samples are tested for deviation from the corresponding theoretical extreme,
2. a "large-sample" test, if $\ln n \ge 10^1$ and deviation can jointly be tested for the first ($\ln n$) extreme values within the the same sample.

The two tests are described, for instance, in and *Glänzel* and *Schubert* (1988) and *Schubert* and *Telcs* (1989). We just mention the following important property. If the distribution of the random variables of the sample is *Paretian*, that is, if the common distribution asymptotically obeys *Zipf's Law* (i.e., $P(X_k = k) \sim c \cdot k^{-\alpha}$ for $\alpha > 0$ and $k \gg 1$), then the determination of the modified Gumbel values is not necessary. In particular, we have

$$P(k(\ln X_k^* - \ln X_{k-1}^*) < x) = P(k \cdot \ln (X_k^*/X_{k-1}^*) < x) \sim 1 - e^{-\alpha x} \; ; \; k \leq k_0 \ll n.$$

Then on the basis of the maximum deviation

$$D_{k_0}(x) = \max | F_{k_0}(x) + e^{-\alpha x} - 1| \text{ with } F_{k_0}(x) = k_0^{-1} \cdot \Sigma \, \chi(k \cdot \ln (X_k^*/X_{k-1}^*) < x),$$

a simple *Kolmogov-Smirnov test* can be applied.

*Other distribution models for scientific productivity*

The application of distribution models to empirical data has shown that the Lotka model could not properly describe different situations of publication activity. The missing ability to describe temporary inactivity, decreasing skewness in growing publication periods as well as different publication habits in the individual science fields may serve just as examples. Practice has shown that an appropriate distribution model needs at least one or two free parameters. The Waring distribution described above is one model meeting these requirements. However, the most versatile model has been introduced by *Sichel* (1985, 1992). In this context we mention that both the above-mentioned *Waring* and *negative-binomial* distribution introduced as results of *birth processes* can be analogously to Sichel's approach be derived from *Poisson distributions* through mixture with continuous distributions. For instance, we could say that scientific productivity of an author (of a given age and social status active in a given field) has a Poisson distribution with parameter $\lambda$ but the parameter $\lambda$ itself is random variable as, for instance, age and status of authors may differ. If $\lambda$ has a *Gamma-distribution* than a *negative-binomial* distribution will be obtain by mixture of the Poisson with the Gamma distribution. A further mixture with a *Pareto-istribution* results in a *Waring-distribution.*

The *generalised inverse Gaussian-Poisson distribution* was originally introduced by Sichel as a comprehensive stationary statistical model for word frequency distribution. Later, he has applied his model successfully to bibliometrics, too. His basic idea is a two-parameter mixture (in the generalised case, a three-parameter mixture) of Poisson distributions obtained by allowing the Poisson parameter $\lambda$ to have a generalisation of an inverse Gaussian distribution. The resulting distribution is very complex (and therefore not presented here); the estimation of parameter is consequently difficult. The goodness of fits exceeds, however, that of all other models.

### *3. 4 Factors influencing publication activity, subject characteristics in publication activity*

Publication activity is influenced by several factors. Separation of factors could result in relative simple models like the Poisson model as described in the previous section. At the micro level, we can distinguish the following four factors.

1. the subject matter
2. the author's age
3. the author's social status
4. the observation period

At higher level of aggregations (e.g., at institutional or national level), the influence of the factors *age* and *social status* vanishes since populations at this level are rather heterogeneous. The other two factors (*subject matter* and *publication period*) can, however, be taken into consideration in the sampling process at the meso and macro level, too.
The publication activity in theoretical fields (e.g., mathematics) and in engineering is lower than in experimental fields or in the life sciences. Experienced authors are expected to be more productive than "newcomers". Publication activity in longer observation periods is obviously greater than in short periods since publication activity is a cumulative process.
An example by Jacobs (2001) illustrates the influence of the status of researchers on their publication activity (see Figure 3.9). Her sample is based on the activity of a selected group of academic and research scientists of ten universities of South Africa for a period of five years, 1992-1996.

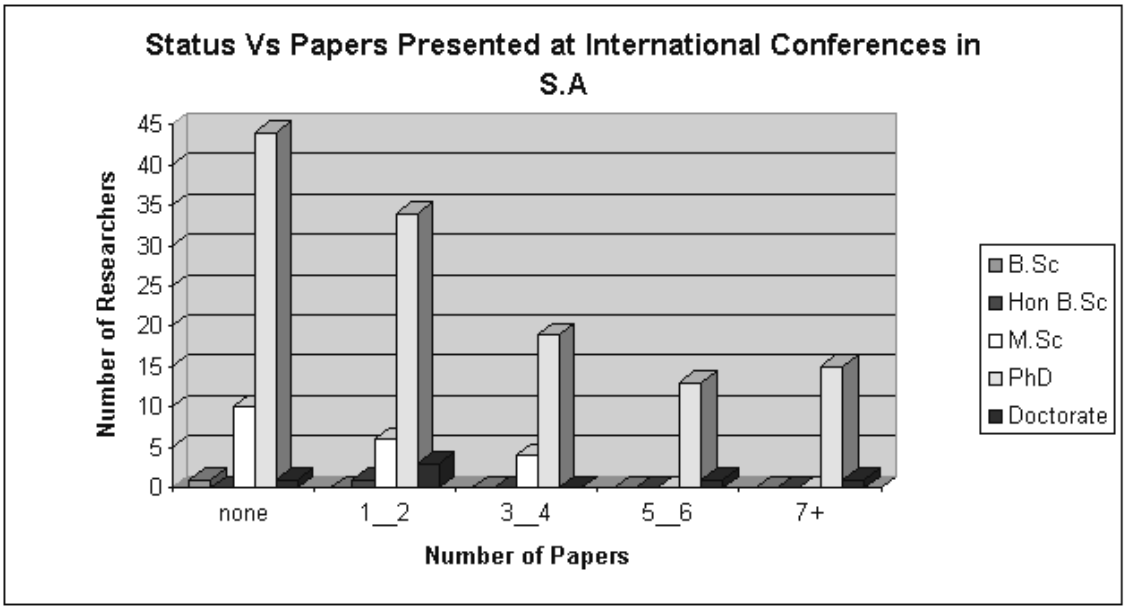



*Figure 3.9 Publication of papers in South African journals (redrawn from Jacobs, 2001)*

*Absolute vs. relative publication indicators*

When publication patters are analysed in bibliometrics, usually *relevant* document types are selected. That is, only those document types that are conveyers of relevant scientific information are taken into consideration. Such publications are, in particular, journal papers of the type *research articles, letters, notes* and *reviews.* Meeting abstracts, editorial material, corrections/errata, retractions, book reviews and other document types not listed above are only objects of special bibliometric studies.

The trend in a unit's publication output is one of the basic measures in research evaluation. Practically two approaches are in use, the change of the absolute number of publications in time and the share of the unit's publication output in that of a higher aggregation. The second approach is applied mainly at the meso and macro level. The reason can be found in the "growth" of the underlying bibliographic databases. Figure 3.10 shows the growth of the SCI during the period 1980-2000. The underlying edition is the CD-ROM version. The number of index documents (of the type *research articles, letters, notes* and *reviews*) in 1980 was

roughly 450,000, the number of journals covered by the database was around 3,250. The growth of documents is almost perfectly linear with correlation coefficient of $r \sim 0.96$. The ”pit” in 1984 is caused by changing profile of the database. This is also reflected by the decrease of journals covered by this edition of the SCI.
Because of the changes in the coverage and the definite growth of the database, the share of the unit's publication output in the total or in a unit representing a higher level of aggregation should be preferred in macro and meso studies.

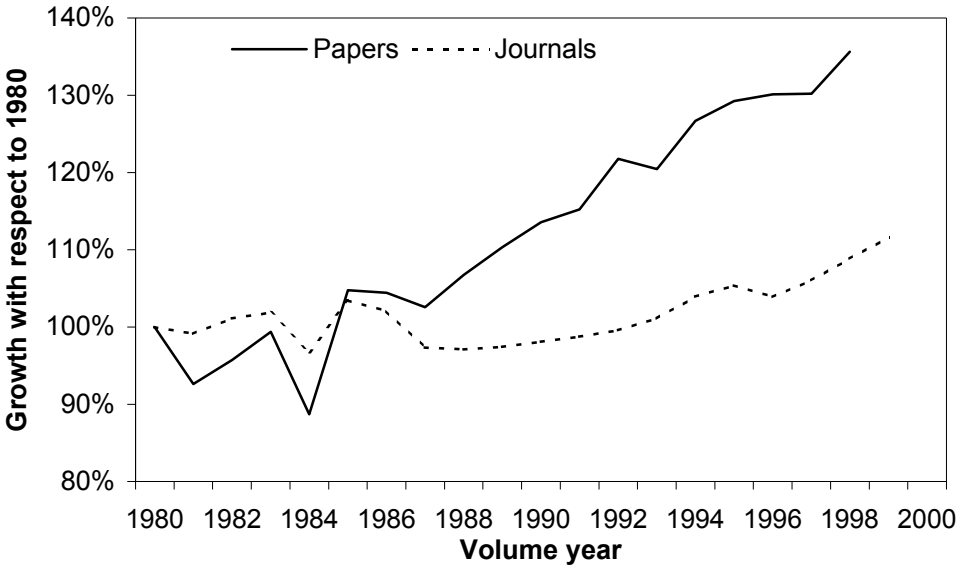



*Figure 3.10 The growth of the SCI database (1980-2000)*

In order to visualise in how far database growth can relativise growth of national publication output, and how dramatic this effect can be, the following example is selected. Figure 3.11 shows the annual change of the Scandinavian share of publications in the world total in two subject fields in the period 1980-1997. The data have been taken from a recent paper by *Glänzel* on Scandinavian science published in 2000. The change of Sweden's productivity is most interesting in this context. The Swedish total in the two selected fields, *chemistry* and *biomedical research* has grown from 1980 to 1997 by 74% and 29%, respectively. However, Sweden's share in the world total grew only in chemistry but decreased clearly in the field of biomedical research. Results of this paper and related studies lead to the conclusion that bibliometric indicators reflect inflationary tendencies that are only partially conditioned by structural changes and growth of underlying bibliographic databases. Within the framework of an ongoing project, *Persson* et al. (2003) have analysed in how far changing patterns of documented scientific communication caused by intra-scientific factors in the last two decades are responsible for such inflationary values. According to their study based on SCI data (CD-Edition), the number papers have grown by 36 percent between 1980 and 1998 that of the authors has increased by 64 percent. This discrepancy in growth patterns cannot be explained by the mere growth of documents indexed by the underlying database. The fact that the growth rate is faster in the case of authors than that of publications forms a contrast to the observed increasing productivity of authors, too. The average productivity increased in the period 1980-2000 by 22% where the share of authors with low productivity decreased. These effects are caused by structural changes in scientific communication, and go much beyond the policy of database producer in preparing their products. In an earlier study, *Kretschmer* and

*Rousseau* (2001) have already analysed the breakdown of Lotka's Law caused by 'author inflation'.

Implementations of the above results for research evaluation are two-fold. The most important consequence of growth of the value of the basic indicators is the need for relative and strictly normalised indicators in bibliometric trend analyses and medium-term or long-term studies. Otherwise, these indicators might show a growth where the development is actually characterised by stagnation or even by decline.

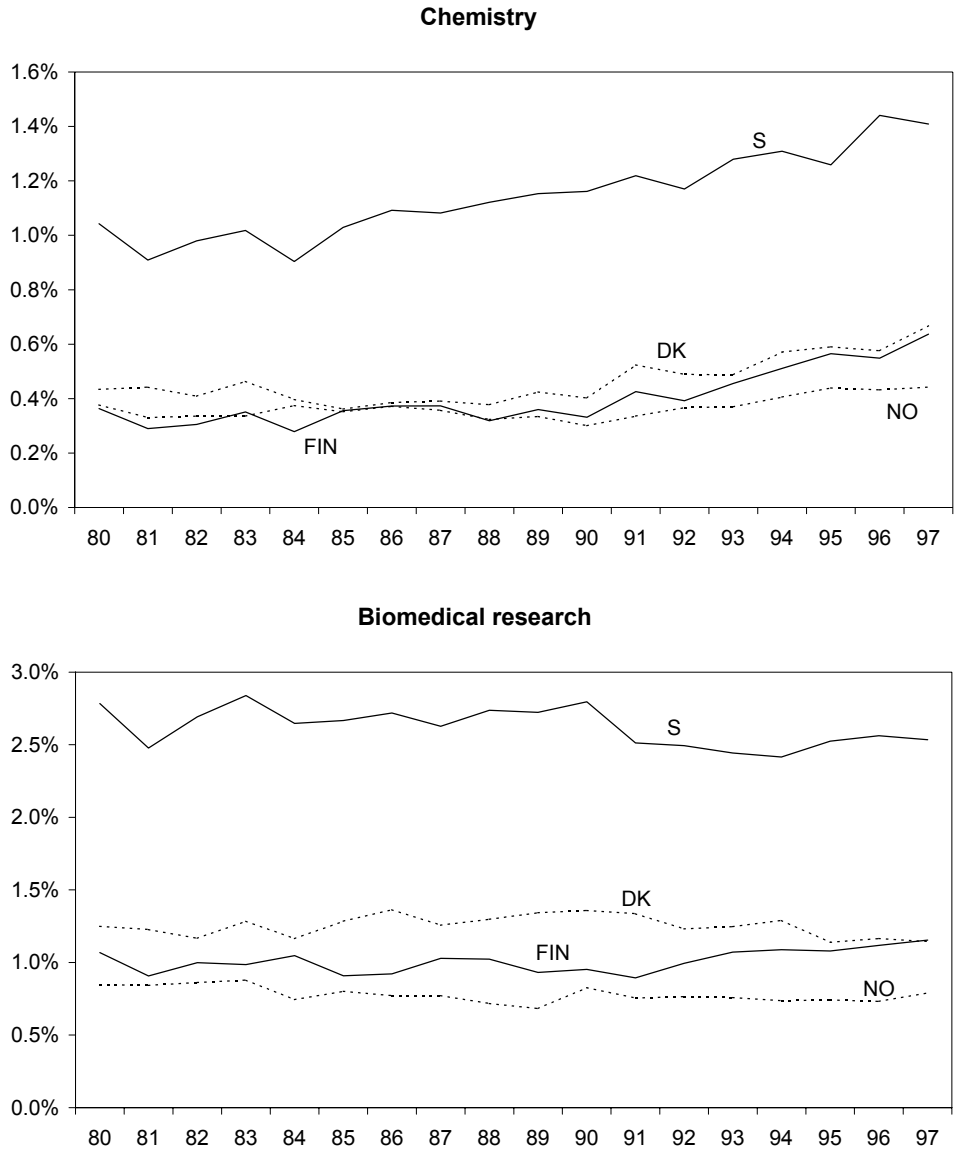



*Figure 3.11 Annual change of Scandinavian share of publications in the world total (1980-1997)*

### *3.5 Publication profiles of institutional and national research activity*

Besides trends in and shares of publication activity, there is a second important issue in bibliometric research of scientific productivity at the meso and macro level. This issue is the analysis of *publication profiles.* Profile can be determined from different perspectives, for

instance, by subject fields (specialisation), by corporate addresses (sectors) and by funding of research.

### *3.5.1 Publication profiles by discipline*

Two important indicators of specialisation are in use, the share of publications of a given unit (institution, region or country) in given science areas in the unit's total publication output and a relative indicator, the *Activity Index* or the *Relative Specialisation Index*, respectively. In the following, we will define an indicator measuring publication profiles with respect to subject fields at the national level, but this indicator can readily be redefined for institutional profiles within a region or country. In particular, it indicates whether a country has a relatively higher or lower share in world publications in a particular field of science than its overall share in world total publications. The *Relative Specialisation Index* (RSI) is closely related to the so-called *Activity Index* (AI) long used in bibliometrics (*Frame*, 1977, *Schubert* and al., 1989), which, in turn, is a version of the economists' Comparative Advantage Index. Activity Index is defined as follows.

$$AI = \frac{\text{the world share of the given country (region) in publications in the given field}}{\text{the overall world share of the given country (region) in publications}}$$

or, equivalently,

$$AI = \frac{\text{the share of the given field in the publications of the given country (region)}}{\text{the share of the given field in the world total of publications}}$$

The Relative Specialisation Index is then defined as

$$RSI = \frac{AI - 1}{AI + 1}.$$

It is easy to see that *RSI* may take its values in the range [-1,1]. *RSI* = -1 indicates a completely idle research field, *RSI* = 1 if no other than the given field is active. *RSI* < 0 indicates a lower-than-average, *RSI* > 0 a higher-than-average activity; *RSI* = 0 reflects a completely balanced, the "average" situation. It is important to note that *RSI* reflects a certain internal balance among the fields in the given unit, that is, positive *RSI* values must always be balanced by negative ones: in no unit can all *RSI* values be positive (or negative).
Four *basic paradigmatic patterns* in publication profiles can be distinguished:

I. the 'western model', that is, the characteristic pattern of the developed Western countries with clinical medicine and biomedical research as dominating fields,
II. the characteristic pattern of the former socialist countries, present Economies in Transition and China with excessive activity in chemistry and physics
III. the 'bio-environmental model', that is, the pattern most typical for developing and more 'natural' countries (e.g., Australia, or South Africa) with biology and earth and space sciences in the main focus,
IV. the 'Japanese model', now also typical for other developed Asian economies with engineering and chemistry being predominant.

Figure 3.12 presents the *Relative Specialisation Index* of two Scandinavian countries in 1987 and 1997 based on eight major fields of science, particularly, *Clinical medicine* (MED), *Biomedical research* (BRE), *Biology* (BIO), *Chemistry* (CHE), *Physics* (PHY), *Mathematics* (MAT), *Engineering* (ENG) and *Earth and space sciences* (ESS). Both Scandinavian countries correspond to Type I, however, Norway's profile can be considered a mixture of types I and III changing more and more to type III. Denmark is changing from a country with almost extremely predominant activity in life sciences to a country representing a more balanced type I profile (cf. *Glänzel*, 2000).

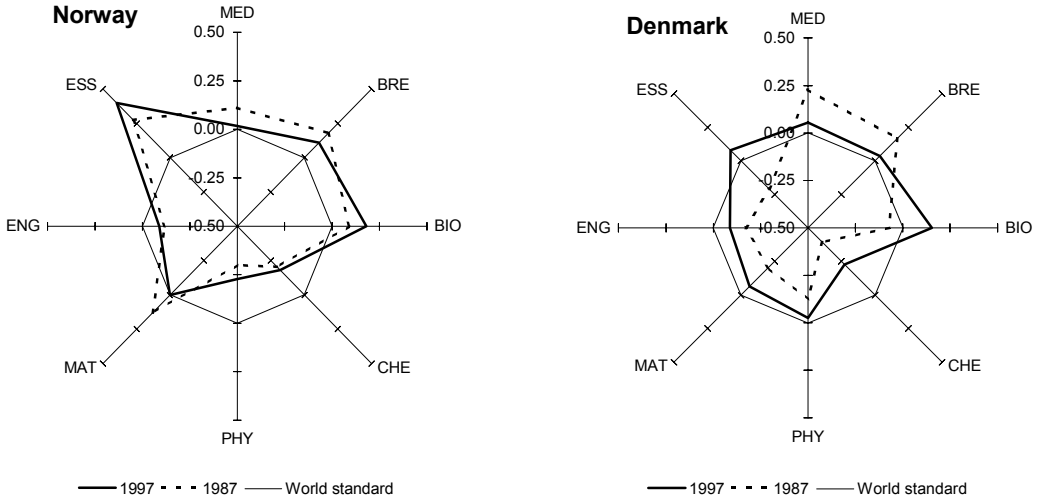



*Figure 3.12 Relative Specialisation Index of the two Scandinavian countries based on eight major science fields (1987 and 1997)*

Finally we will also given an example for type II. The following figure (Figure 3.13) has been taken from the 2nd Edition of the *European Report on Science and Technology Indicators* (REIST-2, 1997). It presents the relative specialisation profile of Romania and Poland in the two periods 1984-1989 and 1990-1995. The predominance of natural sciences in these countries is unusually clear.

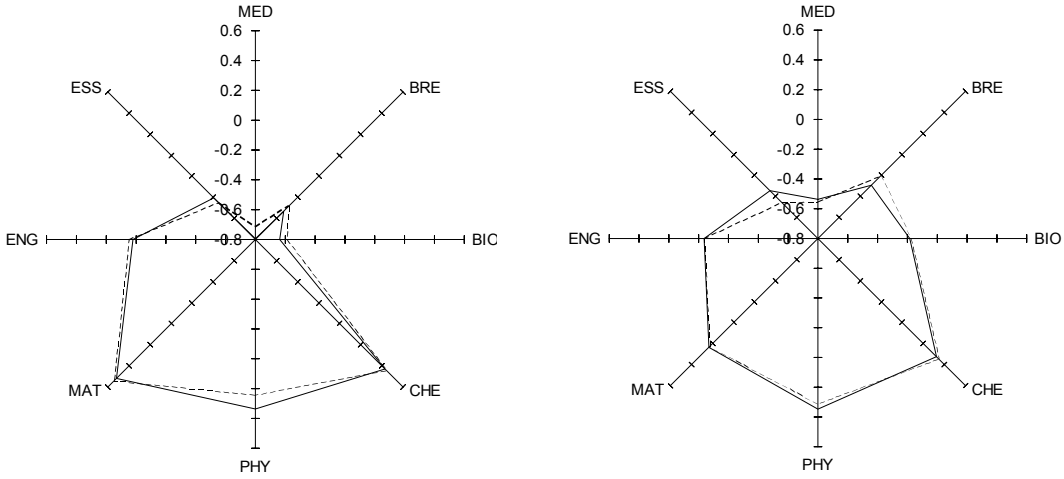



*Figure 3.12 Relative Specialisation Index of Romania and Poland*
*(dotted line: 1984-1989 and solid line: 1990-1995)*

### *3.5.2 Publication profiles by sectors*

Further distinction can be made, for instance, between academic and industrial research, or, within the academic sector, between university and non-university research. The analysis of publication activity in different sectors is a rather delicate question since industry research is less visible through publications scientific journals than academic research. Moreover, many publications are of mixed type (e.g., industry/academic or university/non-university academic research) or the identification of independent or associated institutions proved to be difficult. The profile of Flemish publications by sectors shown in Figure 3.13 has been taken from the latest edition of the "Flemish Indicator book" (*Debackere*, 2003). The underlying publication period was 1992-2001. Flanders has a typical Western-type profile with predominant research at universities and other institutions of higher education. The share of public and governmental research is about one order lower. These data are contrasted by those found for Hungary in the same period. The share of papers published by authors affiliated with universities or other institutions of higher education amounts to roughly 60%. By contrast, the share of public is about 40% that, in turn, is dominated by the *Hungarian Academy of Sciences* with nearly one third of all national publications. The share of independent hospitals with 2% is comparably small. It must be stressed that in both examples the shares cannot be summed up to 100% since there is an extensive collaboration between sectors, in Hungary, above all between the Academy of Sciences and the universities.

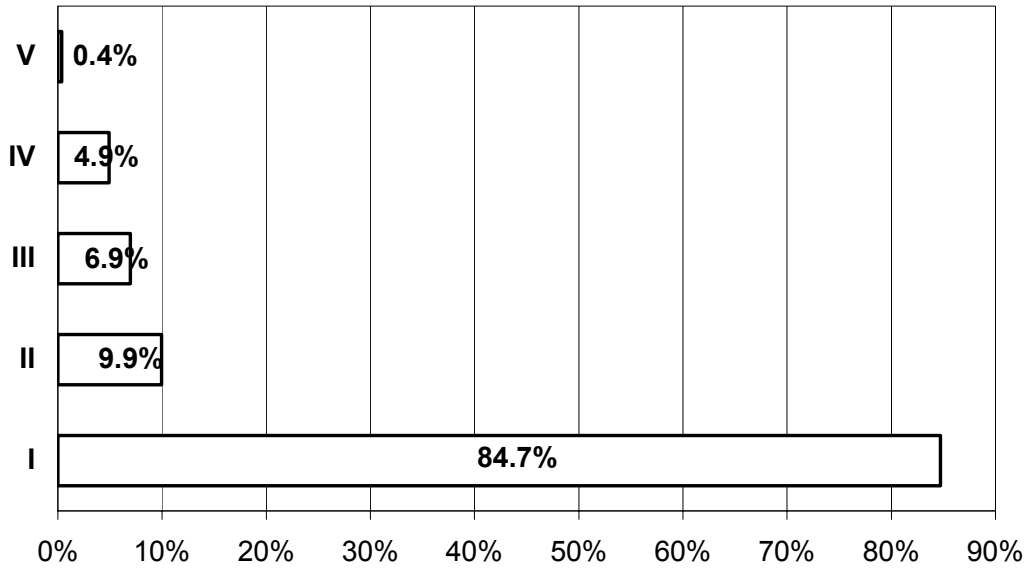



*Figure 3.12 Distribution of Flemish SCI-publications by sectors (1992-2001)*
*I higher education, II public institution or government, III private institution, IV hospital (not associated with universities), V others*

### *3.5.3 Publication profiles by funding*

Aspects of funding can be uncovered through the analysis of acknowledgement in scientific publications. *Grant Lewison* was a pioneer in analysing the use of funding acknowledgement data as a key factor in determining research 'quality' (e.g., *Lewison* and *Dawson*, 1998). He

has studied the behaviour of authors in acknowledging their funding (*Lewison* et al., 1995). They found that a considerable share of authors do not acknowledge funding. Although main sources of governmental, public, institutional or industrial funding can be identified, their extent, however, cannot be estimated reliably on the basis of acknowledgements alone.

### *3.5.4 Characterising research dynamics of institutions, regions or countries*

In order to analyse the research dynamics in selected geopolitical regions, the fields *author* and *address* of the SCI database have been used. Figure 3.13 presents the plot of the indicator *renewal* vs. *transience* in a fixed 5-year publication period according to *Glänzel* (1992b). *Transients* are authors that terminate their publication activity after having published one single paper (cf. Section 2.4.2). The share of transients in the total *scientific community* (i.e., in all authors) is called *Transience Index*. By contrast, the *Renewal Ratio* expresses the ratio of *Newcomers* and not-transient *Terminators*. Figure 3.13 visualises the dynamics of scientific communities in selected countries.

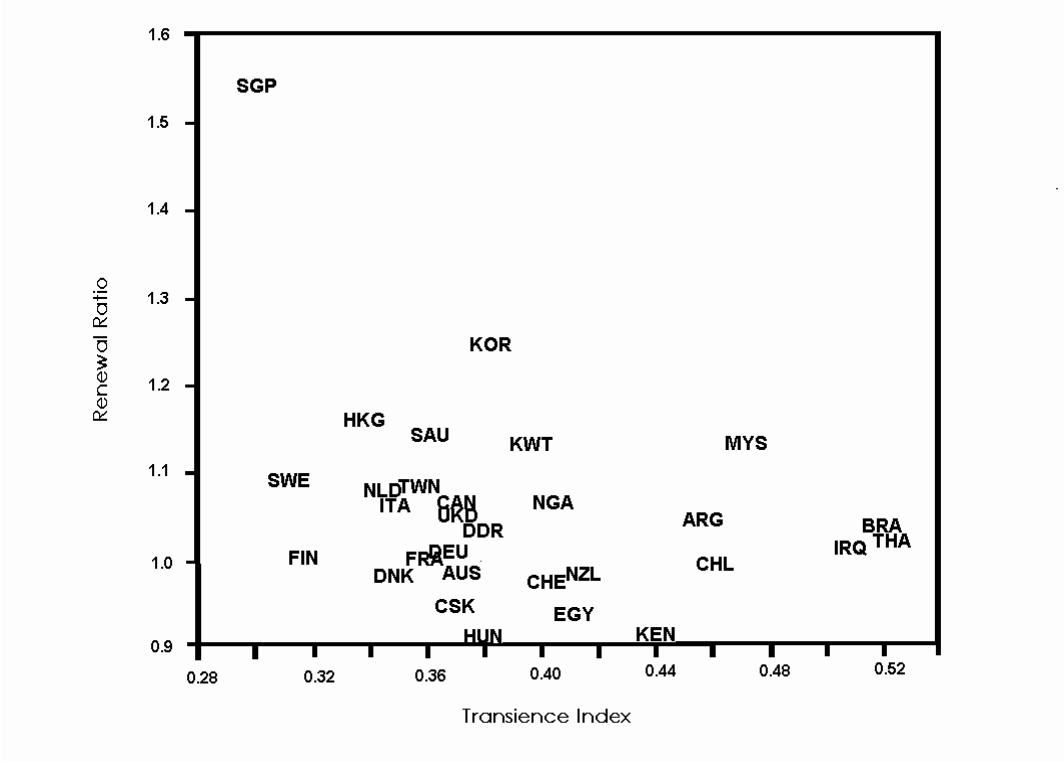



*Figure 3.13 Graphic presentation of the pair of indicators Renewal-Transience for 30 selected countries in 10 geopolitical regions (according to Glänzel, 1992b)*

# 4. INDICATORS OF CITATION IMPACT

## *4.1 The notion of citations in information science and bibliometrics*

In research evaluation, citations became a widely used measure of the impact of scientific publications. There is controversial discussion about how *citations should be interpreted. Susan Cozzens* (1989) has argued that citation is only secondarily a reward system. Primarily, it is rhetorical-part of persuasively arguing for the knowledge claims of the citing document. *Linda C. Smith*, stated that

> *"citations are signposts left behind after information has been utilized".*

*Blaise Cronin* defined citations as

> *"frozen footprints in the landscape of scholarly achievement ... which bear witness to the passage of ideas"*,

but he also referred to certain problems with regard to reference practices as he concluded,

> *"If authors can be educated as to the informational role of citations and encouraged to be more restrained and selective in their referencing habits, then it should be possible to arrive at a greater consistency in referencing practice generally."*

Problems with citation analysis as a reliable instrument of measurement and evaluation have been acknowledged throughout the literature. *Chapman*, for instance, delineated 25 shortcomings, biases, deficiencies, and limitations of citation analysis. *Wouters* (1997) has devoted a large monograph on *citation culture* and in 1998, *Leydesdorff* has initiated the discussion about reappraisal of existing theories of citation.

According to *Westney* (1998), citations are nevertheless indicators of scholarly impact:

> *"Despite its flaws, citation analysis has demonstrated its reliability and usefulness as a tool for ranking and evaluating scholars and their publications. No other methodology permits such precise identification of the individuals who have influenced thought, theory, and practice in world science and technology."*

In recent studies, *Glänzel* and *Schoepflin* (1999) have citations more pragmatically interpreted as

> "*one important form of use of scientific information within the framework of documented science communication*,"

however, in general, whether form of nor reason for the concrete information use are not specified. Although citations cannot describe the totality of the reception process, they give, according to *Glänzel* and *Schoepflin*,

> *"a formalised account of the information use and can be taken as a strong indicator of reception at this level."*

This statement, though made from the perspective of information science, is completely in keeping with the above evaluation-related conclusion drawn by *Westney*.

As mentioned above by *Glänzel* and *Schoepflin*, whether form of nor reason for the concrete information use are taken into account in evaluative studies. However, there are many reasons for citing publications. *Garfield* and *Weinstock* have listed 15 different reasons for giving citations to others' work (cf., *Weinstock*, 1971).

1. Paying homage to pioneers
2. Giving credit for related work (homage to peer)
3. Identifying methodology, equipment, etc.
4. Providing background reading
5. Correcting one's own work
6. Correcting the work of others
7. Criticising previous work
8. Substantiating claims
9. Alerting to forthcoming work
10. Providing leads to poorly disseminated, poorly indexed, or uncited work
11. Authenticating data and classes of facts – physical constants, etc.
12. Identifying original publications in which an idea or concept was discussed
13. Identifying original publications or other work describing an eponymic concept or term
14. Disclaiming work or ideas of others (negative claim)
15. Disputing priority claims of others (negative homage)

This list is, of course, not exhausting, but some of the above reasons for being cited, for instance, # 5-7, # 14 and # 15, may illustrate that not all given citations point to quality. But even criticism expresses the reception of documented scientific information. Heavy criticism of a certain scientific work can, in a sense, reflect true impact. By provoking constructive criticism, an erroneous theory may even more contribute to the advancement of a science area than some sound average study. On the other hand, papers of a controversial nature will continue to be cited longer.

These examples may just serve as an illustration of the complexity of citation processes. The general discussion of sociological theories of citations is beyond the scope of this introduction to citation indicators. For further arguments and a detailed presentation of citation contexts we, therefore, refer besides the already cited work to the articles by *Small* (1978, 1982) and *Bonzi* and *Snyder* (1991).

These reasons listed by *Garfield* and *Weinstock* can be categorised, on one hand, as 'positive', 'neutral' and 'negative' and, on the other hand, as relevant, less relevant and even irrelevant or redundant. Figure 4.1 visualises the 'weight' of citations.

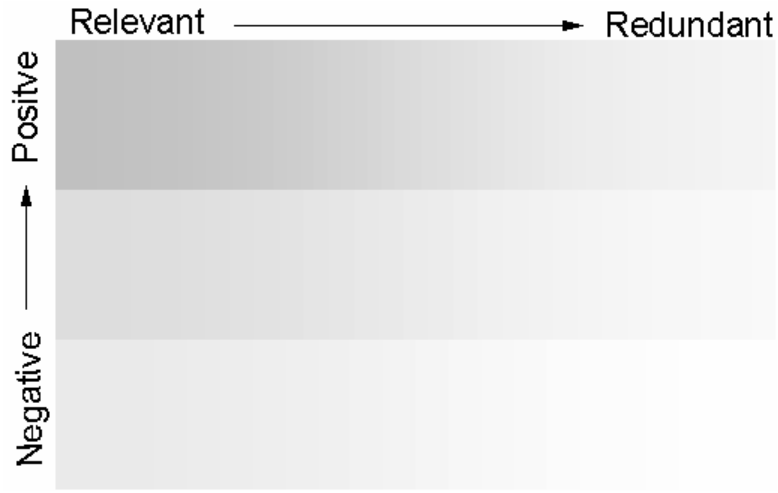



*Figure 4.1 The 'weight' of citations from the viewpoint of the use of scientific information*

The question arises in how far citations express the use of information and the reasons according to *Garfield* and *Weinstock* in reality. Nevertheless, the refereeing system in documented scientific communication guarantees the observance of *relatively strict rules* of providing reference citations.

It might also be worthwhile to list reasons for not giving citations to a colleague's work, that is, for *not* providing reference citations. The first and most important one is *lacking relevance* of the topic. Irrelevant topics are obviously not cited. *Unawareness* is the second reason that is due to insufficient retrieval of published information relevant for an author's research work. Citations omitted by reason of unawareness are sometimes added by referees reviewing papers prior to acceptance for publication in a journal. According to *Garfield* (1986) these papers not recognised by unawareness are whether

> *"victims of cryptomnesia, an unconscious plagiarism in which creative ideas expressed as new are actually unrecalled memories of another's idea, or were superseded for other reasons remains to be seen."*

*Disregard* has little do with *bibliographic amnesia* described by Garfield. Disregard is simply a reason that is already beyond the borderline to unethical communication behaviour. Results by colleagues relevant for the author's research to be published are this way demonstratively ignored. The fourth reason is a consequence of *obsolescence* as an expression of 'natural' *obliteration*. Finally, the fifth reason occurs rather seldom in scientific literature, as it is an expression of the *disappearance* of the 'users' of information. In order words, literature still being relevant in the context of the research topic is not cited because there are no more authors who could cite it. We can consider such topics *extinct.*

As already mentioned in the 2. Section, ageing and obsolescence is expressed by changing frequency of citations (received or given) over time. So, what is actually measured by decreasing citation impact? First and above all, decreasing citation impact reflects, of course, obsolescence. This might be an *evolutionary* process, that is, cited work gradually vanishes from reference lists, and is replaced by more recent and more relevant literature. However, it might be a *revolutionary* process if a breakthrough has at once made everything obsolete that

was so far relevant in this subject. A third form of obliteration occurs if literature is subject to *obliteration by incorporation*. This means relevant literature is no longer cited because its substance has been absorbed by current knowledge; its content has thus become "common knowledge" (*Garfield*, 1977, 1986). *Garfield* (1997) concluded that

> *"obliteration ... is one of the highest compliments the community of scientists can pay to their author".*

The author or inventor has become *eponymous*. Examples for such eponymic concept or term in bibliometrics are Bradford's Law and Lotka distribution.

In this context it should be mentioned that all these arguments and reasons refer to *individual* citations. Bibliometrics, however, deals in most cases with citations to larger sets of publications. As in case of other social processes, a difference in the behaviour of individuals and the behaviour of large groups or masses can be detected in citation processes, too. Critics of bibliometric methods argue that citations might be governed by the will and the (sometimes tendentious) intentions of the authors and may thus result in a deliberate filtering of information sources (e.g., "citation cliques"). Individual citing may be influenced hereby, but this phenomenon is certainly not characteristic for the citations to paper sets published by a larger number of authors. Thus, especially reasons # 10, # 14 and # 15 of Weinstock's list are less characteristic in citations to large publication sets. Citation analyses are based largely upon citation frequency. Several authors in bibliometrics are downright regarding citation frequencies as a "quality measure". This interpretation is not only narrowing down the possibilities of application of citation-based methods, it may also have undesired consequences at the micro-level, if publications of individuals are studied.
For instance, the fact that a paper is less frequently cited or (still) uncited several years after publication gives information about its reception by colleagues but does not reveal anything about its quality or the standing of its author(s). Uncited papers by Nobel Prize winners may just serve as an example. However,

> "*if a paper receives 5 or 10 citations a year throughout several years after its publication, it is very likely that its content will become integrated into the body of knowledge of the respective subject field; if, on the other hand, no reference is made at all to the paper during 5 to 10 years after publication, it is likely that the results involved do not contribute essentially to the contemporary scientific paradigm system of the subject field in question*" (*Braun* et al., 1985).

### *4.2 The role of self-citations*

*Self-citations* are a special type of citations: Several forms of self-citations can be distinguished; two of them are of special importance: Author self-citations and journal self-citation. Both forms have to be clearly distinguished from each other. Journal self-citation occurs if a paper published in a given journal is cited by a paper published in the same journal. A great share of journal self-citations allows the conclusion that the journal in question is highly specialised, a low share indicates in a sense a "lack of originality"; a low share of journal self-citations (for instance, < 10%) is, for example, characteristic for review journals (see, *Schubert* and *Braun*, 1993). Journal self-citations are also interesting in the context of obsolescence studies. Ageing of journal self-citations can significantly differ from

that of "foreign" publications. Figure 4.2 (redrawn from *Glänzel* and *Schoepflin*, 1995) illustrates this effect.

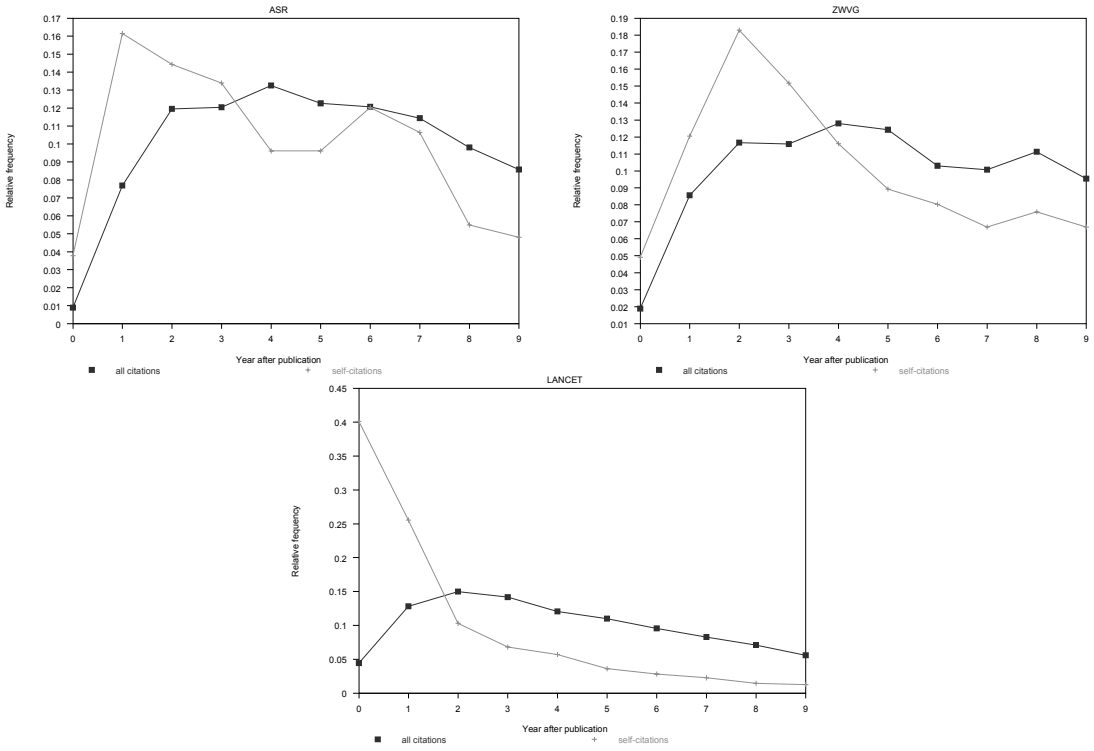



*Figure 4.2 Distribution of citations (black) and (journal) self-citations (grey) over time (American Sociological Review, Probability Theory and Related fields, The Lancet)*

Author self-citation occurs if an author refers to a own paper, that is, if he was the author or one of the co-authors of the cited paper. The spectrum of these self-citations ranges from the obvious case, if an author refers to his own work, to more hidden forms, if the co-authors of the author in question are citing him or themselves in another paper.
Several bibliometricians are inclined to omit at least the obvious types of author self-citations. In the evaluative context at the micro-level this practice seems to be justified (see *Bonzi* and *Snyder*, 1991, *Nederhof* et al, 1993). Above all, users in science policy, but sometimes even the researchers themselves are almost "condemning" author self-citations as possible means of artificially inflating citation rates and thus of strengthening the authors' own position in the scientific community. Other Bibliometricians are rather inclined to regard a reasonable share of author self-citations as a natural part of scientific communication. According to this view, it is quite normal that a scientist or a research group refers to the own work. Thus self-citations do not reveal much about the true impact of research. The share of self-citations in all citations as well as their share in all references reveals interesting aspects of an author's or a research group's role in the system of science communication. The almost absolute lack of self-citations over a longer period is just as pathological as an always-overwhelming share. The first one may indicate lack of originality in research, whilst the latter symptom indicates isolation and lacking communication. The great number of self-citations – provided, of course, the share in all citations exceeds not the normal extent – also indicates a successful and dynamic publication activity since the author or group has then published numerous papers in refereed journals.

*MacRoberts* and *MacRoberts* have given a first overview of the unsolved problem of self-citations in their critical review on problems of citation analysis in 1989.

Beside the discussion on the principles of the role of author-self citations, there is no real consensus concerning how this type of self-citations should be defined operatively. In practice, two different approaches to *direct* self-citations are in use. At the micro level, that is, on the level of individual authors, a self-citation for an author *A* occurs whenever *A* is also (co-)author of a paper citing a publication by *A*. This definition cannot, however, be applied to higher levels of aggregation, that is, when publications and citations are aggregated over sets of *different* (co-)authors, and the notion of self-citations is uncoupled from an individual author *A*. The definition of self-citations suggested by *Snyder* and *Bonzi* (1998) and *Aksnes* (2002) can preferably be used at these levels of aggregation. According to this method, a self-citation occurs whenever *the set of co-authors of the citing paper and that of the cited one are not disjoint*, that is, if these sets share at least one author. Although the reliability of this methodology is affected by homonyms (resulting in Type II errors by erroneous self-citation counting) and spelling variances/misspellings of author names (resulting in Type I errors by not recognising self-citation), at high levels of aggregation, that is at the meso and macro level, there is no feasible alternative to this method.

A recent large-scale analysis of author self-citations by *Glänzel* et al. (2003) gives interesting insight into the mechanism of scientific communication.
The first important result of this study characterises to the relationship between self-citations and foreign citations. Self-citations and foreign citations proved not to be independent variable. Moreover, the conditional expectation of self-citations for given number of foreign citation could be characterised by a *square-root law*. This shows that there is from the statistical viewpoint nothing arbitrary in self-citations. Thus self-citations are an essential part of scientific communication. Figure 4.3 visualises this square-root law for author self-citations.

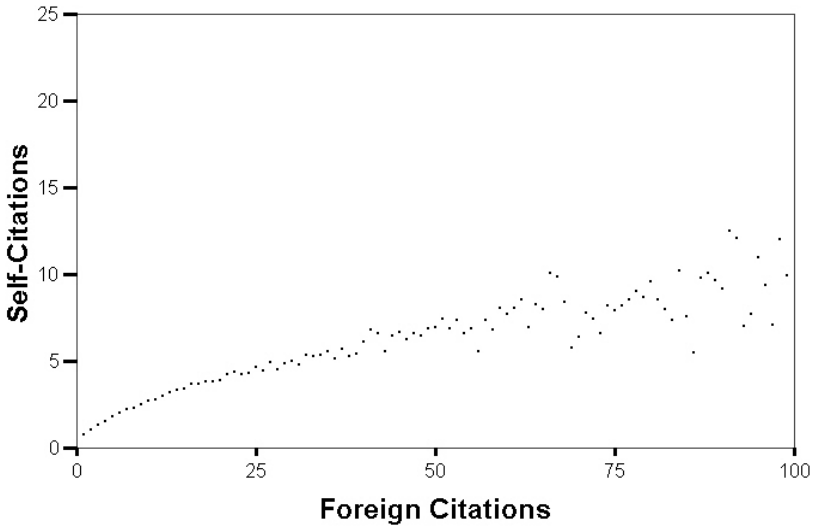



*Figure 4.3 Square-root law for author self-citations*
*(Plot of expected number of self-citations over number of foreign citations)*

Analogously to journal self-citations, the weight of self-citations decreases rapidly. The process is even faster in the case of author self-citations: In the third year after publication, the distribution of expected self-citations over foreign citations is practically stationary.
Figure 4.4 presents the distribution of author self-citations and foreign citations (i.e., non-self citations) over time in all fields combined. The figure has been redrawn from *Glänzel* et al. (2003).

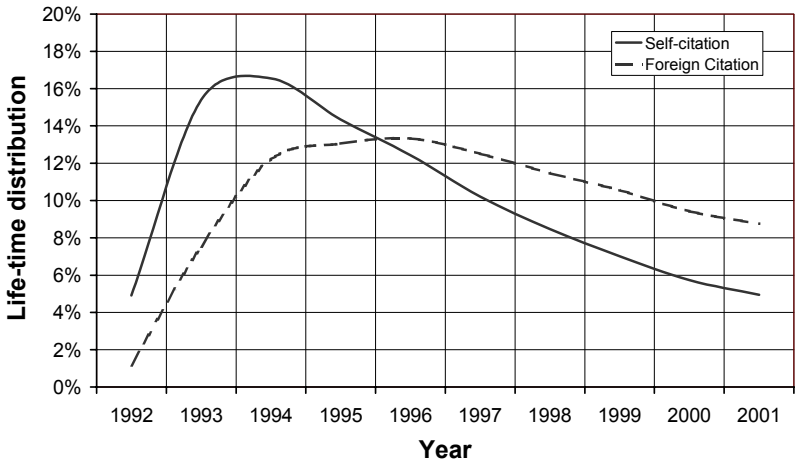



*Figure 4.4 Distribution of author self-citations and foreign citations over time (all fields combined)*

The above study has also shown that *low visibility* goes with *high self-citation shares*. Figure 4.4 visualises this effect.

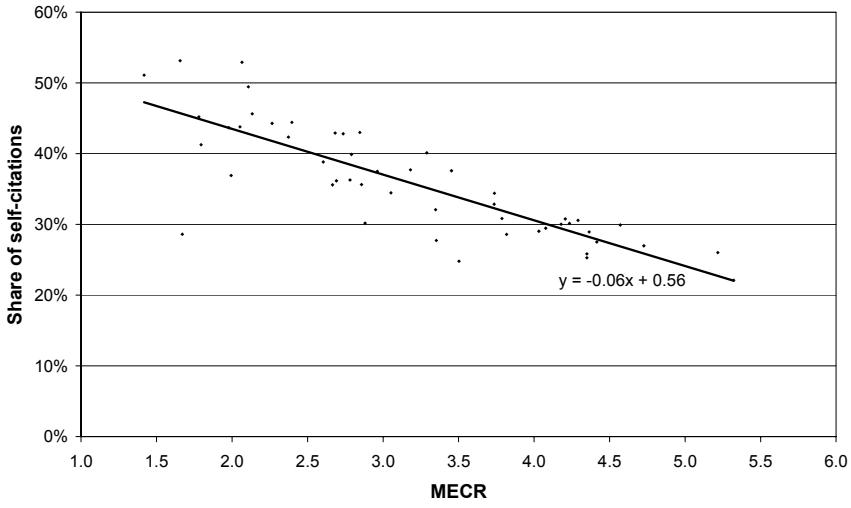



*Figure 4.4 Plot of self-citations vs. visibility according to Glänzel et al. (2003)*

Indicators based on author self-citation proved to be valuable supplementary measures that can be used both in informetrics and research evaluation. Because of restrictions concerning their reliability, self-citation indicators should be used in addition to traditional citation indicators, but not replace them. There is no reason for condemning self-citations in general. At the level of individual authors or of research teams, the deviation of the share of

self-citations from the subject-specific standard might be taken into consideration in evaluative application of bibliometrics.

### *4.3 Factors influencing citation impact*

The reference items of scientific publications reflect characteristics concerning the "hardness" of scientific literature. Price (1970) used the share of references not older than five years in all references of a journal to distinguish between hard science, soft science, technology, and non-science (*Price Index*). *Moed* redefined this index in 1989 as share of most recent references for individual papers (*Price Index per paper*). In a large-scale study of 1981, *Line* analysed structure of social science literature, and investigated what makes social science different. Recently, *Egghe* (1997) and *Glänzel* and *Schoepflin* (2001) have published further studies on these topics.

Another indicator for the distinction between "hard" and "soft" science has been found by *Glänzel* and *Schoepflin* (1999). The percentage of references to serials characterises typical differences in the communication behaviour in the sciences, social sciences and humanities (see Figure 4.5).

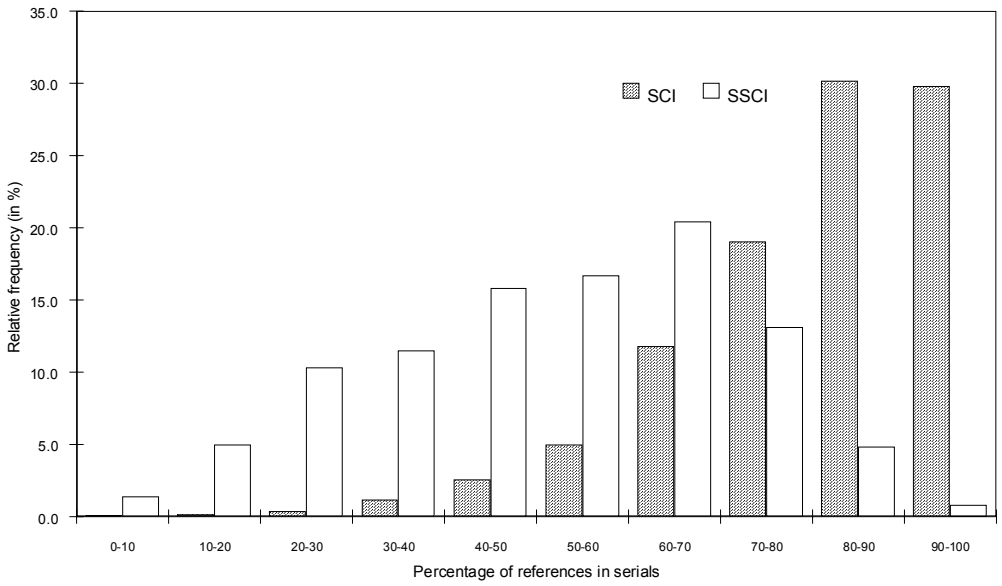



*Figure 4.5 Distribution of the share of references to serials over journals in the sciences and social sciences*

The above question, namely what share of references to serials and non-serials, respectively, is typical of "hard" and " soft" sciences can be extended in the following manner. The more general question arises which forms of information "sources" and "targets" play a role in science communication. In the social sciences and, even more, in the humanities a considerable part of cited information is originated in non-science literature. In case of engineering, the target is in part outside the scientific community; information is used, e.g., for the advancement in technology.

| | |
|---|---|
| **Authors:** | Mirza-K Michael-A |
| **Title:** | Major Depression in Children and Adolescents |
| **Full source:** | BRITISH JOURNAL OF HOSPITAL MEDICINE 1996, Vol 55, Iss 1-2, pp 57-61 |
| **Language:** | English |
| **Document type:** | Review |
| **IDS/Book No.:** | TV978 |
| **No. Related Records:** | 20 |
| **No. cited references:** | 19 |
| **Cited references:** | ANGOLD-A-1995-DEPRESSED-CHILD-ADOL-P127<br>GOODYER-I-1985-BRIT-J-PSYCHIAT-V147-P517<br>GOODYER-IM-1993-J-CHILD-PSYCHOL-PSYC-V34-P1409<br>HARRINGTON-R-1990-ARCH-GEN-PSYCHIAT-V47-P465<br>HARRINGTON-RD-1993-DEPRESSIVE-DISORDER-P1<br>JENSEN-PS-1992-J-CHILD-ADOLESC-PSYC-V2-P31<br>KOVACS-M-1984-ARCH-GEN-PSYCHIAT-V41-P643<br>KOVACS-M-1986-DEPRESSION-YOUNG-PEO-P435<br>KOVACS-M-1993-J-AM-ACAD-CHILD-PSY-V32-P8<br>KOVACS-M-1995-DEPRESSED-CHILD-ADOL-P281<br>KUTCHER-SP-1989-PSYCHIAT-CLIN-N-AM-V12-P895<br>KUTCHER-SP-1995-DEPRESSED-CHILD-ADOL-P195<br>LEFKOWITZ-MM-1978-PSYCHOL-BULL-V85-P716<br>**MANN-T-1929-DISORDER-EARLY-SORRO-P26**<br>PFEFFER-CR-1993-J-AM-ACAD-CHILD-PSY-V32-P106<br>RIE-HE-1966-J-AM-ACAD-CHILD-PSY-V5-P653<br>ROCHLIN-G-1959-J-AM-PSYCHOANAL-ASS-V7-P299<br>STROBER-M-1992-J-CHILD-ADOL-PSYCHOP-V2-P23<br>WEISSMAN-MM-1987-ARCH-GEN-PSYCHIAT-V44-P747 |

*Figure 4.6 Example for cited information originated in non-science literature in case of Psychology/Psychiatry*

The interpretation of the concept of citation as *one important form of use of scientific information within the framework of documented science communication* according to *Glänzel* and *Schoepflin* (1999) does not contradict the application of citation-based indicators to research evaluation studies, since frequently (or rarely) used information disseminated, say, by the scientific community of a country or institute is certainly symptomatic for the research performance of the community in question. Citation based measures are preferably designed for use in the assessment of research in the natural sciences, the life sciences and mathematics. They can, however, be applied with certain restrictions to the bibliometrics of engineering and selected fields of the social sciences.

Citation impact is mainly influenced by the following five factors that are analogously to the case of publication activity at higher levels of aggregation practically quite inseparable.

1. the subject matter and within the subject, the "level of abstraction"
2. the paper's age
3. the paper's "social status" (through the author(s) and the journal)
4. the document type
5. the observation period

*Subject Characteristics of Citation Based Indicators*

Citation patterns are strongly influenced by subject characteristics. Citation measures are therefore – without normalisation – not appropriate for cross-field comparisons. Citation measures for different multidisciplinary papers sets might be distorted by the underlying

publication profiles. The following example might illustrate this effect (source year: 1996, citation window: 1996-1998).

| *Subject field* | *Mean citation rate* |
|---|---|
| Mechanical, civil and other engineering | 1.12 |
| Mathematics | 1.46 |
| Analytical chemistry | 3.00 |
| Solid state physics | 3.06 |
| Neurosciences | 4.54 |

In 1983, *Peritz* presented a study on the intra-disciplinary differences in citation impact of theoretical, methodological, and empirical papers in sociology in three prestigious sociology journals. In natural sciences, theoretical subjects have usually lower impact than applied ones.

*The paper's "social status" and the observation period*

The following example (see Figure 4.7) shows the influence of "social status" (through the journal) and observation period (ranging between one year and 21 years) on citation impact. The cumulative impact of the prestigious journal of the American Chemical Society *Analytical Chemistry*, the German journal *Fresenius Journal of Analytical Chemistry* published by Springer-Verlag and the formar Soviet *Journal of Analytical Chemistry of the USSR* is compared.

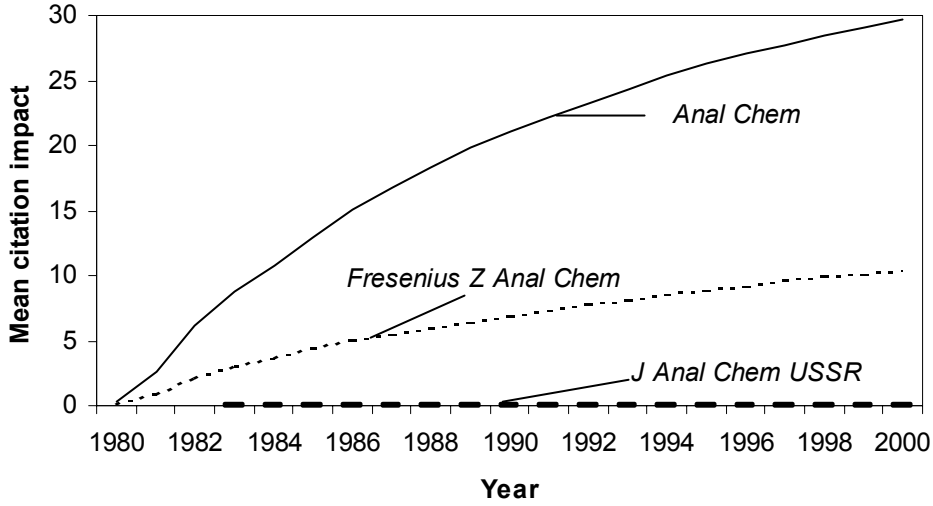



*Figure 4.7 Example the influence of "social status" and observation period on citation impact*

The third example illustrates the influence of a complex of several factors on citation impact. The impact factor of two journals *American Sociological Review* (ASR) and *The Lancet* are compared in dependence of time. Besides subject peculiarities here the document type is responsible for the deviating trends of the journals. A considerable part (more than 60%) of all documents were *Letters* in 1980.

| Citation window | Mean citation rate ASR | LANCET |
|---|---|---|
| **1980-80** | 0.2 | 0.6 |
| **1980-81** | 1.8 | 2.4 |
| **1980-82** | 4.3 | 4.5 |
| **1980-85** | 12.1 | 9.7 |
| **1980-89** | 20.9 | 14.0 |

*Figure 4.7 Mean citation rates of two journals in time as a function of time visualising the influence of several factors on citation impact (publication year: 1980)*

### *4.4 Journal Citation Measures: The Impact Factor*

Journal citation measures are one of the most widely used bibliometric tools. They are used in information retrieval, scientific information, library science and research evaluation. And they are applied at all levels of aggregation.
The main source of journal citation measures is the annually appearing *Journal Citation Report* (JCR). The most important measure is the *Impact Factor* (*Garfield*, 1979). The impact factor for the journal $J$ in the year $n$ is defined as the ratio

$$IF_n(J) = \frac{c_n}{p_{n-1}+p_{n-2}},$$

where $c_n$ is the number of citations received in the year $n$ by papers published in the journal $J$ in the years $n$–1 and $n$–2 and the total number of source items ($p_{n-1}+p_{n-2}$) published in the journal $J$ in these two years ($n$–1 and $n$–2).

The *Immediacy Index* is defined analogously to the Impact Factor as a journal citation measure of citations received in the publication year, particularly,

$$II_n(J) = \frac{c_n}{p_n}.$$

The strengths of the *Impact Factor* lies above all in its independence of the "size" of the journal, its comprehensibility, stability and seeming reproducibility. On the other hand, some obvious flaws, but especially the already mentioned uninformed use have provoked critical and controversial discussions about its correctness and use. In this context, it has also to be mentioned that ISI's somewhat poor background documentation concerning the processing of the data presented in the JCR cannot convince critical users. In particular, the IF and related journal impact measures can readily be reproduced from the data presented in the Journal Citation Reports (JCR), however, these very data proved at large not to be reproducible. Although it is difficult to theoretically define the concept of (journal) impact, there is a wide spread belief that the ISI *Impact Factor* is affected or 'disturbed' by factors that have nothing to do with (journal) impact. Consequently, several attempts have been made to improve the impact factor or to develop additional or alternative journal citation measures. Some of the main modifications relate to all of the 'elements' mentioned in the above-mentioned mathematical interpretation.

- Instead of the mean: other parameters of the distribution (e.g. percentage of uncited papers or quantiles)
- Instead of integer counting of citations: weighting a citation on the basis of the journals in which it is made
- Instead of applying a single citing year: application of a range of citing years
- Instead of analysing all ('citable') documents: disaggregate articles on the basis of document type (article + note + review) or content (e.g., theoretical, methodological and experimental)
- Instead of considering only papers 1-2 years earlier: analysing articles from older 'ages'
- Instead of synchronic: diachronic, or a combination of the two approaches

Because of its comprehensibility, robustness and its fast availability, the impact factor became very quickly popular and widely used. The Impact Factor is *comprehensible* because it measures the frequency with which an average article published in a given journal has been cited in a particular year; it is *robust* because the annual changes of the journals' Impact Factors proved to be not dramatic so that in practice one or two years old impact factors are sometimes used for evaluation purposes where more recent indicators are not available. On the other hand, time series can be used to monitor the evolution of journals' citation patterns. The *fast availability* of the Impact Factor, finally, is due to the fast indexing, data processing and the distribution of ISI products. These are in short the most important technical advantages of the Journal Impact Factor.
On the other hand, according to a number of authors both the Impact Factor and especially the Immediacy Impact have several serious flaws the consequences of which shall be discussed here.

1. There is no normalisation for reference practices and traditions in the different fields and disciplines (*Pinski* and *Narin*, 1976).
2. "There is no distinction in regard to the nature and merits of the citing journals" (*Tomer*, 1986).
3. There is a bias in favour of journals with large papers, e.g. review journals tend to have higher impact factors (*Pinski* and *Narin*, 1976).
4. Citation frequency is subject to age bias (*Asai*, 1981, *Rousseau*, 1988, *Glänzel* and *Schoepflin*, 1995, *Moed* et al., 1998).
5. There is no indication of the deviations from this statistic (see, for instance, *Schubert* and *Glänzel*, 1983).
6. The average time for a journal article from publication to peak in citations is not always two years, or as *Garfield* (1986b) writes "if we change the two-year based period used to calculate impact, some type of journals are found to have higher impacts". (cf. also *Glänzel* and *Schoepflin*, 1995, *Moed* et al., 1998)
7. One single measure might not be sufficient to describe citation patterns of scientific journals.
8. The concept of citable document is not operationalised adequately. As a result, journal impact factors published in ISI's Journal Citation Reports are inaccurate for a number of journals (*Moed* and *van Leeuwen*, 1995, 1996).

9. In the calculation of JCR impact factors, errors are made due to incorrect identification of (cited) journals, for instance for the journal *Angewandte Chemie* – International Edition (*Braun* and *Glänzel*, 1995, *van Leeuwen* et al, 1997).

The above-mentioned limitations lead very early to the discussion of possible improvements or alternatives. Among others, *Yanovski* (1981) criticised certain distortions of the impact factor and suggested new indicators based on the ratio between citations and references. Yanovski's indicator has not found wider use although, for instance, *Smart* and *Elton* (1982) and *Todorov* (1983) have critically reacted on his approach. Thus, *Smart* and *Elton* showed that the consumption factor and the impact factor are statistically independent which suggests that these two measures represent distinct journal attributes.

In 1978 *Lindsey* introduced the *Corrected Quality Ratio* defined as (number of citations)$^{3/2}$/(number of publications)$^{1/2}$. This formula can be reformulated as the product of the square root of the impact factor and the number of citations. This approach, however, lacks interpretability. This might be one reason why this indicator has not found application.

Allison has given an interesting but undeservedly neglected approach in 1980. He used the statistical function (standard deviation – mean value)/(mean value)$^2$ as an inequality measure of distributions of scientific productivity and citation impact. The underlying assumption is that the distribution of authors by the number of publications or by the frequency of citations is *negative binomial*. This is the first time that a particular distribution model is assumed for citation frequency. The indicator is the reciprocal of an estimator of one of the parameters $N$ of the distribution.

*Schubert* and *Glänzel* (1983) have studied the statistical reliability of journal *Impact Factors*. In particular, they analysed both the significance of the deviation between the impact factors of the same journals calculated by different institutes and of the deviation between the impact factors of two different journals representing the same discipline (see Section 2.2.4). They have used a similar model as suggested by Allison. The results of this study have strong methodological and practical influence on journal rankings by impact factors and comparative analyses in research evaluation.

According to *Asai* (1981) the period count based on a month produces more accurate statistics than that based on a year. The author introduces an *Adjusted Impact Factor* which counts the weighted sum of citations over a period of four years instead of one year as in case of the original *Impact Factor*.

The most sophisticated improvement has been presented by *Narin* and *Pinski*. Whereas in calculating the impact factor and the immediacy index all citations are equally weighted, the "influence methodology" suggested by *Pinski* and *Narin* (1976) provides for each journal a size-indepedent *Influence Weight* determined by the number of the journal's citations and references. The calculation is based on an iteration procedure involving great expense. Weighting citations by these influence weights, the *influence per publications* and the *total influence* can be calculated. *Geller* (1978) suggests a 'corrected' influence weight that could be interpreted as the probability that a given journal will be cited from the other journals. Because of the troublesome calculations and the lack of expressive interpretability of the results, the method has gained few adherents.

In the last two decades, several bibliometric research centres therefore succeeded in calculating their own journal impact measures on the basis of the bibliographic databases of the ISI. Since 1995 citations are often counted in a three-year or four-year observation period: in the year of publication and the two subsequent years at the ISSRU in Hungary and at RASCI in Germany. This three-year citation window proved to be a good compromise between the relatively fast obsolescence of technology oriented literature, of most areas in life sciences, of experimental physics literature, on one hand, and of the slowly ageing theoretical and mathematical topics in physics, on the other (see, *Glänzel* and *Schoepflin*, 1995 and *Moed* et al., 1998). CWTS has also used four-year observation period in several studies. An overview of applications, problems and limitations of the Impact Factor has been given by *Glänzel* and *Moed* (2002).

### *4.5. Towards relative citation indicators*

The Citation Rate per Publication (*Mean Observed Citation Rate*) is a fundamental citation based measure that can be applied to all levels of aggregation. MOCR is called *Citations per Publications* (*CPP*) at CWTS. The impact factor discussed in the previous section is a special case of a mean observed citation rate.

*Peter Vinkler* is using mean citation rates in combination with journal impact factors for institutional evaluation of research performance (micro and meso level). He has defined several indicators on the basis of journal impact, field impact and observed citation rates. By averaging *weighted* indicators and constructing relative indices, Vinkler created complex measures designed for application to the evaluation of research performance of individuals and departments (for instance, *Vinkler*, 2002).

The Expected Citation Rate per Publication (*Mean Expected Citation Rate*). The expected citation rate of a single paper is defined as the average citation rate of all papers published in the same journal in the same year.
Note that these journal averages are not necessarily identical with the Journal Impact Factors, as they are defined and listed in the Journal Citation Report volumes of the SCI (*Garfield*, 1975). Any appropriate citation window can be used instead of one year citation window to publications of the two preceding years as used in the *Journal Citation Report* (JCR).
For a set of papers assigned to a given unit (e.g., institution, country or region) in a given field (subfield) the indicator is the average of the individual expected citation rates over the whole set. At CWTS this indicator is called *Mean Citation Rate of Journal Packet* (*JCSm*).

The mean citation rate of a subfield can be considered a second expectation of a paper published in this discipline. At CWTS, this measure is denoted by *FCSm.*

In order to overcome the shortcomings caused by subject characteristics, and to find a fair basis for comparisons, citation measures can be normalised by a proper reference standard. The impact factor of the journal in which the papers have been published and/or the subfield to which they can be assigned may serve as such reference standard. *Relative citation indicators are measures which gauge observed citation impact against the expectation or against a proper reference standard.*

- The *Relative Citation Rate* (*RCR*) is defined as the ratio of the Citation Rate per Publication to the Expected Citation Rate per Publication, that is,

*RCR = MOCR/MECR* (see, for instance, *Schubert* and al., 1989). A version of this measure, particularly, *CPP/JCSm* is used at CWTS.

Both MOCR *and* MECR have to be determined for exactly the same publication year and the same citation window!

RCR measures whether the publications of a unit attract more or less citations than expected on the basis of the average citation rates of the journals in which they appeared. Since the citation rates of the papers are gauged against the standards set by the specific journals, it is largely insensitive to the big differences between the citation practices of the different science fields and subfields. Therefore, this indicator is uniquely suitable for cross-field comparisons.

– The *Relative Citation Impact Index* (RCII) is closely related to the RCR. Similarly to the way *RSI* was derived from *AI*, *RCR* can be normalised into the range [-1,1] by the transformation

$$RCII = \frac{RCR - 1}{RCR + 1}.$$

*RCII* = -1 corresponds to uncitedness, *RCII* = 1 to the fictitious case of infinite number of citations; RCII < 0 means lower-then-average, *RCII* > 0 higher-than-average citation rate, *RCII* = 0 if the set of papers in question attracts just the number of citations expected on the basis of the average citation rate of the publishing journals.

- The *Normalised Mean Citation Rate* (NMCR) is defined as the ratio of the Citation Rate per Publication to the weighted average of the mean citation rates of subfields. A similar measure (*CPP/FCSm*) is used at CWTS.

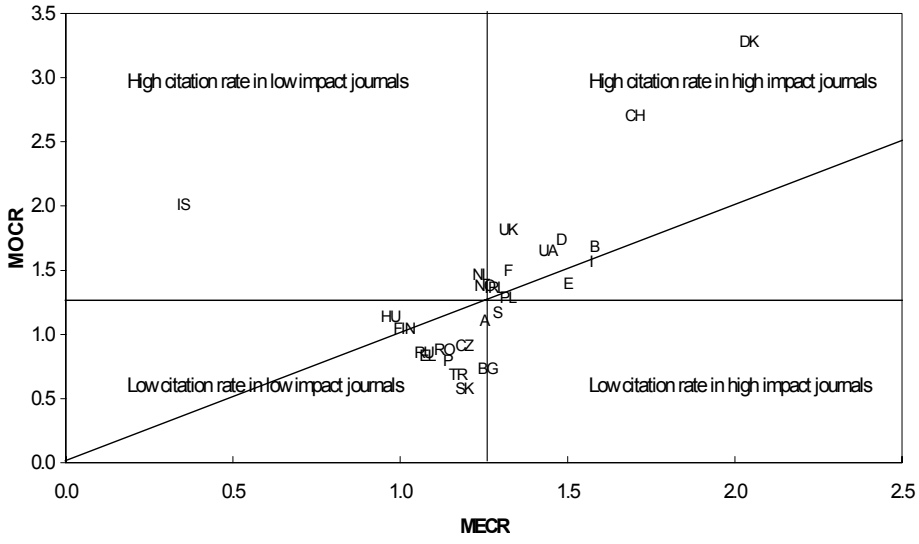



*Figure 4.8 Comparison of citation impact of 26 European countries in mathematics (source year: 1993, citation window: 1993-1995)*

The two relative indicators can large deviate from each other. The reason for that lies in the *publication strategy* of the units of analysis. If, for instance, RCR is significantly higher (lower) than NMCR then authors assigned to the unit in question publish in lower(higher)-than-average journals (with respect to the field).

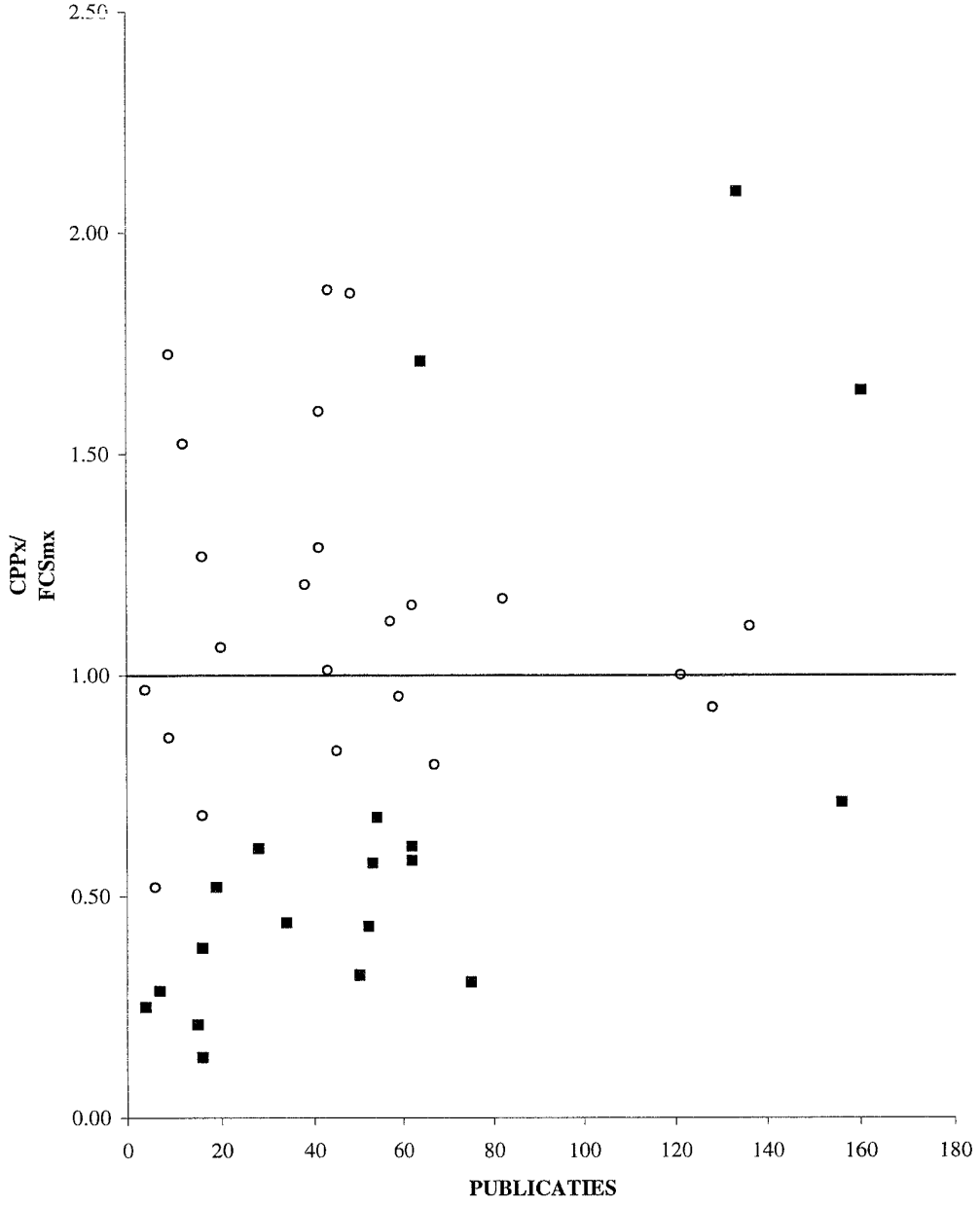



*Figure 4.9 Comparison of the normalised citation impact of departments of the medical faculty of the University Antwerp (1992-1998) using a simple statistical test according to van Leeuwen et al. (2001)**

* Each circle or square represents one department; black coloured squares above (below) the horizontal reference line represent departments for which the impact (*CPP*) is significantlyabove (below) the world average (*FCSm*).

Observed citation based indicators and their expectations can be presented in *relational charts*. Figure 4.8 presents the plot MOCR vs. MECR in mathematics. The indicator values are based on publication data of the 1993 volume of the SCI and the 3-year citation window 19993-1995. According to the definition, MOCR and MECR for subject areas defined on journal basis coincide. The MOCR for mathematics in 1993 amounts to 1.23. Since the selected countries are rather large, the "random error" of MOCR being proportional to $n^{1/2}$, where $n$ is the sample size, is relatively small (cf. Section 2.4.4). However in countries with less than 100 publications, the standard deviation of MOCR is already measurable. The standard deviation of MECR is practically zero since the samples are here based on the underlying journals, that is, is roughly corresponds to that of the total field.

Standard deviation of mean citation impact at the meso level (for instance that of research groups or departments) is usually significantly greater than that at the macro level. Increasing the "random error" by one order has dramatic consequences on conclusions drawn from comparative analyses. The plot in Figure 4.8 shows that a quite large deviation of the mean citation rate of a research group or department from its expectation (or from another one) does not necessarily be considered significant.

*Share of uncited or cited papers*

Although bibliometric indicators applied at a national or supra-national level are decidedly more reliable than their meso and micro level pendants, their use does not lack some typical problems, either. In particular, the higher the level of aggregation the greater the heterogeneity of the population and the statistical distributions underlying the indicators get extremely polarized. Thus, the complexity of the national/supranational citation patterns can scarcely be reflected by a single indicator such as the average citation rate. In other words, one single parameter is not sufficient for describing citation patterns otherwise the expected value would uniquely determine the shape of the distribution. Figure 4.10 might just serve as a counterexample. The two selected journals *Trends in Genetics* and *American Journal of Respiratory and Critical Care Medicine* have almost the same mean citation rate, namely 7.0 and 6.9, respectively (publication period 1995-1996, 3-year citation windows). The shapes of the two distributions is, however, characterised by different features. The share of uncited or cited papers being the estimate of the corresponding probabilities $P(X = 0)$ and $P(X \geq 0) = 1 - P(X = 0)$, respectively, could, of course, serve as an obvious addition measure. *Moed* et al. (1999), however, showed that the two indicators are practically not independent.

The standard error of shares such as the *share of cited papers* ($f_c$) can be calculated similarly to the formula given in Section 2.4.4 for the impact. In particular, we have $D(f_c) = \sqrt{p_c\,(1-p_c)/n}$ , where $p_c$ is the corresponding probability of being cited. The standard error of the share of uncited papers coincides with that of the previous one since $p_0 = 1 - p_c$.

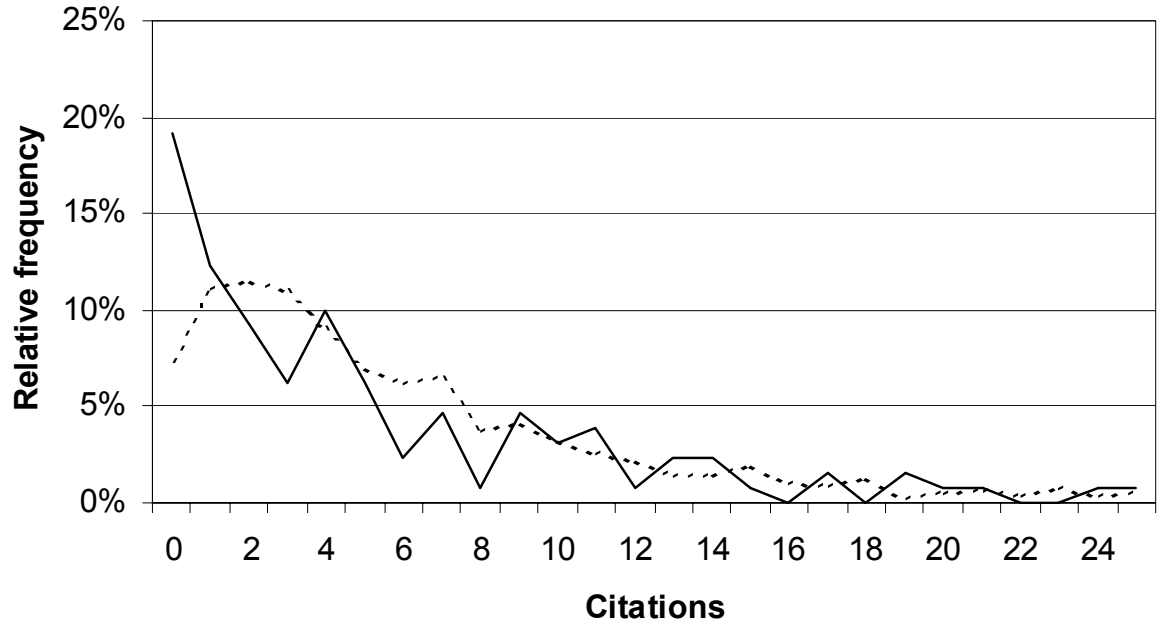



*Figure 4.10 Citation distributions of two journals with same impact and different shape (Trends in Genetics (solid line) and American Journal of Respiratory and Critical Care Medicine (dotted line))*

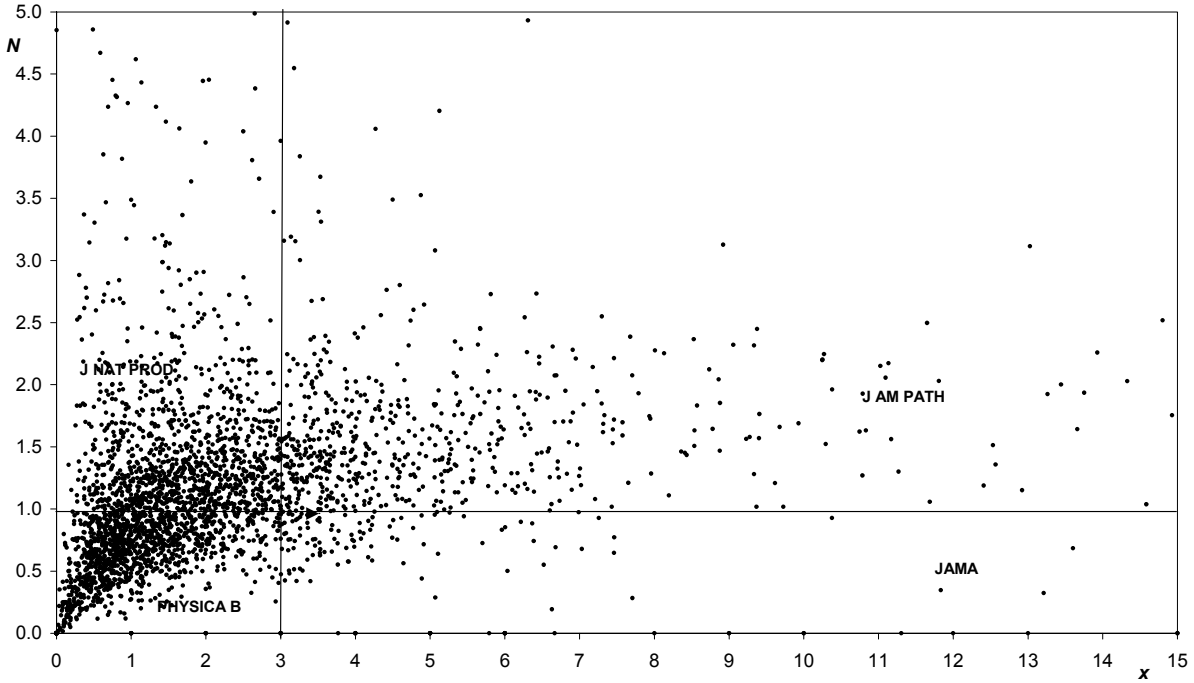



*Figure 4.11 Plot of estimated N values vs. mean x for 3370 journals published in 1995-1996 and cited in 3-year citation windows*

The approach by *Allison* (1980) defines a second parameter that is practically independent of the mean. According to section 4.4, this is the reciprocal of the parameter $N$ in the negative-binomial model by *Glänzel* and *Schubert* (1995). Figure 4.11 presents the plot of estimated parameter $N$ vs. mean $x$ for all journals published in 1995-1996 and cited in 3-year citation windows, each. Except for extremely low impact, there is no significant correlation. Four journals have bee selected to represent the four quadrants (low impact with small and large parameter $N$; high impact with small and large parameter $N$).

*Highly cited publication characetrising the 'high end' of citation impact*

In the last subsection, we have seen that uncited papers may give further information on the degree of polarisation, concerning the 'low end' of the distribution. From the viewpoint of evaluative analyses, characteristics of the 'high end' of the distribution are by far more relevant. Data on *highly cited* papers and authors have regularly been presented (for instance by *Garfield*), although a consensus on the exact definition and meaning of the term highly cited does not exist as yet.
To determine the authors or papers with most citations usually a fixed number (*Vlachý*, 1986) of items or a certain quantile is selected from a rank statistic (for instance, 'the top decile papers' (*Hofer Gee* and *Narin*, 1986)). In some lists, papers or authors are considered highly cited if the number of citations received by them simply exceeds a given fixed value (for instance, "papers cited more than 400 times" (*Garfield*)). These *a priori* criteria reflect neither field-specific peculiarities nor deviations caused by the particular choice of publication and citation periods. At best empirical results may help to find an individual number of items or an individual quantile for each subject field and time period (for example, according to *Garfield*: "in some fields with fewer researchers, 100 citations may qualify a work"). According to *Glänzel* and *Schubert* (1992) thresholds determining highly cited papers should meet the following criteria:

1. They should be great enough to guarantee that the selected items form a real *elite*. On the other hand, it should be small enough to obtain a statistical population of *elite* items.
2. They should be flexible in order to compensate for the unequal publication and citation behaviour in different science fields and to allow "fine tuning" in order to adjust the size of selected groups.
3. The threshold should be time invariant with respect to the citation window.

Thresholds for highly cited papers could be defined as follows. We say that a paper highly cited if the number of citations it has received during a given period exceeds $k_{s(j)} = s \cdot \max(1,x_j)$, where $x_j$ is the average citation rate of the reference standard. The mean citation rate of journal in which the paper has been published or of the subject to which the paper belongs or a combination of the two might be used as the reference standard. The same citation window has to be used for both the citation rate of t he paper and the chosen reference standard.
In verbal terms, a paper is considered highly cited if it has received at least $s$ citations, and the number of citations amounts at least $s$-times the reference standard. The coefficient $s$ is responsible for adjusting the final group size of selected papers. The term $\max(1,x_j)$ contains a fixed component which has two functions, it filters noise and makes sure that the mean citation rate of highly cited papers increases with rising thresholds, i.e., with growing $s$. Despite the advantages of this approach, there is still something arbitrary in the definition of

the coefficient *s*. Moreover, it might occur that a highly cited paper (meeting to the above criteria) does not classify for the selection if a larger citation window is chosen and the threshold has to be redefined for e new reference standard. This makes the "theory" of high citations from the dynamic perspective somewhat instable. However, this applies to almost all solutions discussed here.

# 5. INDICATORS OF SCIENTIFIC COLLABORATION

Scientific collaboration has become one of the favourite topics in bibliometric research. Collaboration pattern can be studied at almost all levels; co-operation of individual scientists has, for instance, been investigated in the context of social stratification in science (e.g., *Kretschmer*, 1992a,b). The published results of intra- and extramural collaboration have been compared at the institutional level, and the strongly intensifying domestic and international collaboration has served as a base of several bibliometric macro studies.

## *5.1 Co-authorship as a measure of scientific collaboration*

Bibliomterics is measuring scientific collaboration by means of co-publication statistics. Of course, the question arises in how far collaboration is reflected by corresponding co-authorship. In a recent study, *Laudel* (2002) has shown on the basis of a sample of interviewed scientists that a major part of collaboration is not acknowledged either through a proper acknowledgement or through co-authorship. However, this rather applies to so-called *intramural* collaboration, that is, to collaboration within a department, research group or institute. Extramural collaboration, above all international collaboration, is, however, well acknowledged. Results of collaborative research at this level are reflected by corresponding co-authorship of the published results that can, in turn, be analysed with the help of bibliometric methods.

Besides the economic and political factors, many intra-scientific factors (see, for example, studies by *deB*. *Beaver* and *Rosen* 1978, 1979, *Luukkonen* et al., 1992, 1993), especially changing communication patterns and increasing mobility of scientists, are also influencing collaboration. These factors motivates co-operation in "less expensive" areas such as pure mathematics and theoretical research in social sciences, too. Analogously to the *Garfield/Weinstock* list of reasons for providing citations, there is also a compilation of motivation for collaborative research. *DeB*. *Beaver* (2001) lists the following 18 purposes for which people collaborate:

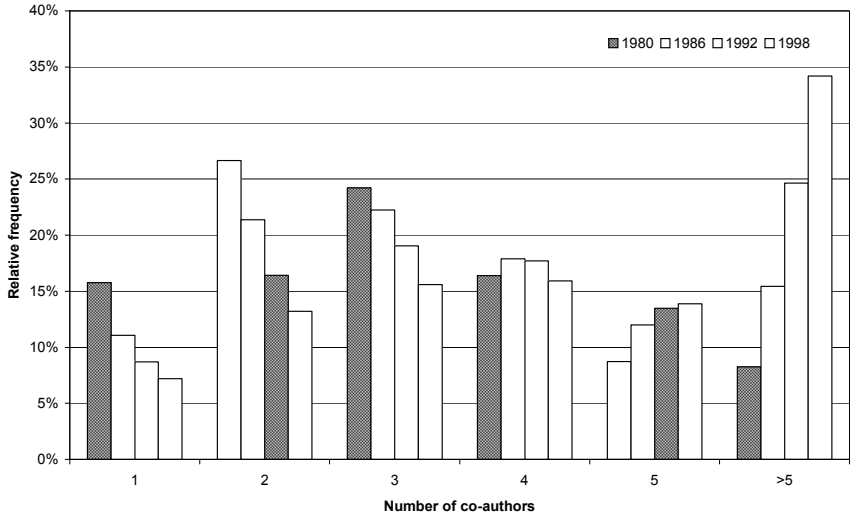
1. Access to expertise.
2. Access to equipment, resources, or “stuff” one doesn’t have.
3. Improve access to funds.
4. To obtain prestige or visibility; for professional advancement.
5. Efficiency: multiplies hands and minds; easier to learn the tacit knowledge that goes with a technique.
6. To make progress more rapidly.
7. To tackle “bigger” problems [more important, more comprehensive, more difficult, global].
8. To enhance productivity.
9. To get to know people, to create a network, like an “invisible college”.
10. To retool, learn new skills or techniques, usually to break into a new field, subfield, or problem.
11. To satisfy curiosity, intellectual interest.
12. To share the excitement of an area with other people.

13. To find flaws more efficiently, reduce errors and mistakes
14. To keep one more focussed on research, because others are counting on one to do so.
15. To reduce isolation, and to recharge one's energy and excitement.
16. To educate [a student, graduate student, or, oneself]
17. To advance knowledge and learning.
18. For fun, amusement, and pleasure.

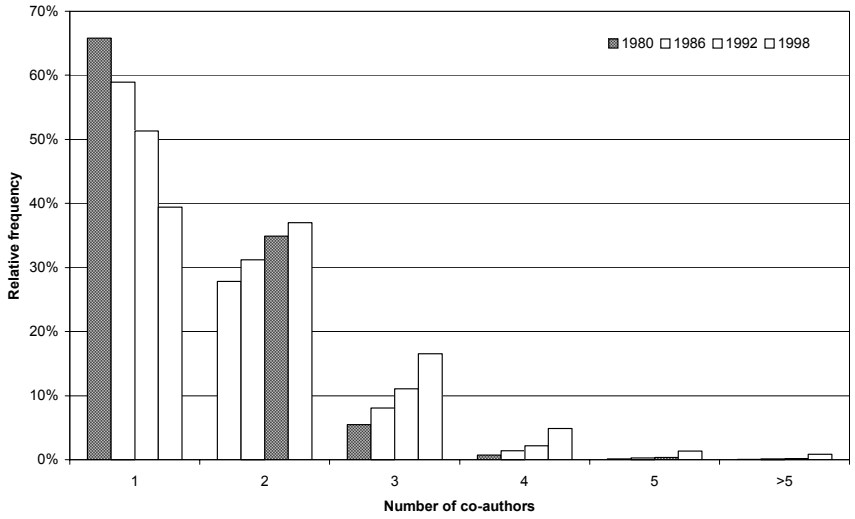
*Kretschmer* (e.g., 1994) has analysed aspects of social stratification in scientific collaboration at the micro (individual) level. Main findings are that extramural collaboration is characterised by similarity of the social status whereas intramural collaboration shows significant differences of the social status of the co-authors.

Scientific collaboration – as measured by means of co-authorship patterns – has considerably increased during the last decades *at all levels of aggregation*. In the 90s the rapid increase has, however, somewhat slowed down.
The most frequently used aggregation levels in the context of co-publication studies are the level of *individual authors*, *extramural domestic*, *collaboration between sectors* (e.g., university/higher education, non-university academic, industry) and *international collaboration.*

The following example shows the increase at the level of individual authors. Figures 5.1a and 5.1b present the distribution of co-authors over publications in 1980, 1986, 1992 and 1996 in two science areas, namely in *Biomedical* research and in *Mathematics* (cf., *Glänzel*, 2002). Figure 5.1b shows that in 1996 already more that 60% of all mathematical papers was multi-authored.




*Figures 5.1a The distribution of co-authors over papers in 'Biomedical research'*




*Figures 5.1b The distribution of co-authors over papers in 'Mathematics'*

Although the increase of collaboration itself is already a remarkable phenomenon, the question arises in how far increasing collaboration interacts with other processes. The assumption that collaboration might increase publication output is almost obvious. If an author is part of a stable team and as such co-author of its publications, he/she shares all papers with the other members of the team and this will be accordingly reflected by integer publication counts. Although this idea suggests itself, it could not be proved empirically. The following example taken from the above by study by *Glänzel* (2002) contradicts any common-place notion. Figure 5.2 presents the plot of average productivity vs. average co-operativity in three selected fields in 1996.

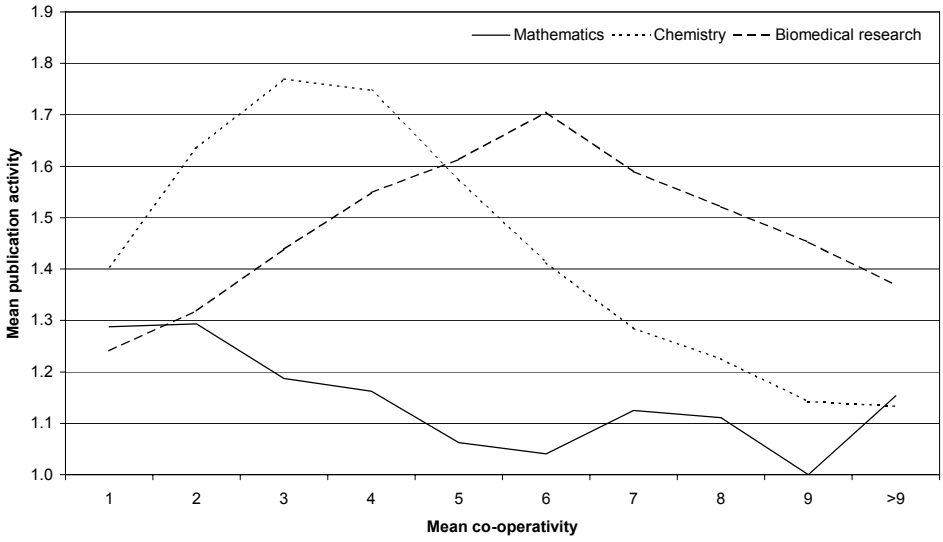



*Figure 5.2 Plot of average productivity vs. average co-operativity in three selected fields in 1996 according to Glänzel (2002)*

For authors in all three fields there is a peak of productivity close to the average co-operativity value characteristic for the field in question. For instance, in biomedical research the maximum productivity is reached for teams with 6 co-authors, whereas in mathematics, mean publication activity takes its maximum value in case of 1-2 co-authors. Otherwise, no unambiguous "effect" on publication activity can be found for the number of

authors involved. Collaboration is thus not associated with higher productivity at the individual authors. In mathematics, productivity is even slightly decreasing with growing co-publication activity. Although "team work" exhibits higher productivity than single authorship in the two other fields, beyond a field-characteristic level productivity distinctly decreases with growing co-operativity. *Braun* et al. (2001) have shown a similar effect in the field of neuroscience.

The positive effect of collaboration and especially of international collaboration on *citation impact* has been shown in many studies, although international collaboration does not always pay for all partners involved (cf. *Glänzel*, 2001). Figure 5.3 visualises this effect by presenting the plot of mean citation rate vs. number of co-authors for domestic and international collaboration for papers published in 1980 and 1998 in all fields combined. For this example, a 3-year citation window has been used.

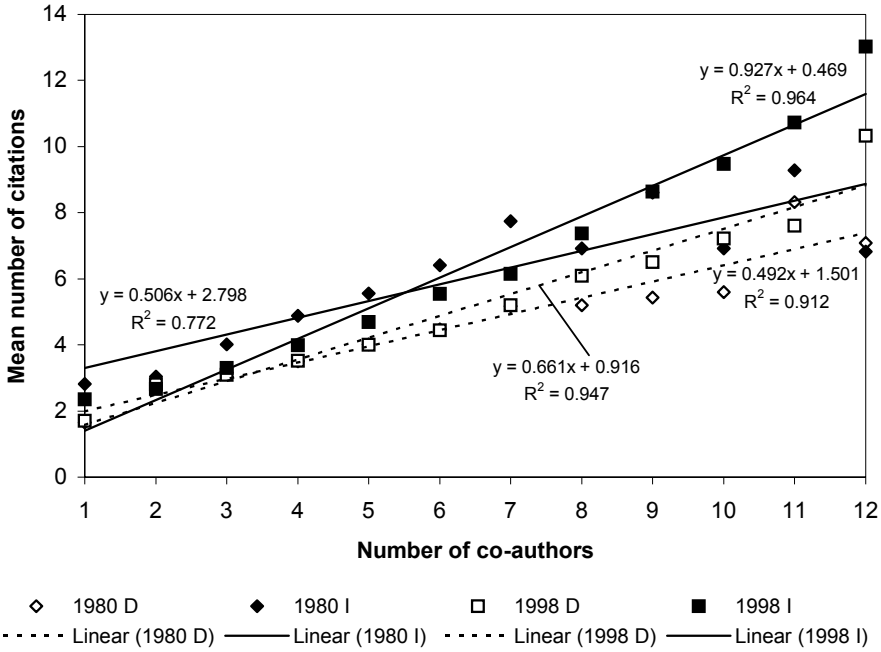



*Figure 5.3 Relationship between number of co-authors and mean citation rate controlling for international collaboration (all fields combined)*

## *5.2 Indicators of co-operativity and co-publication networks*

Most important measures of co-operativity are, for instance, the number and share of co-authored papers of a unit, of joint publications of different units, of the strength of co-authorship links and the profile and citation impact of co-publications. The first comprehensive study on international collaboration using co-operativity measures has been published by *Schubert* and *Braun* in 1990.

Figure 5.4 shows the dramatic increase of international collaboration on the basis of SCI publications. Nevertheless, it should be mentioned that this overall trend does not apply to all countries; the share of internationally co-authored publication, for instance, in Turkey, Korea and Saudi Arabia decreased. Some changes are obviously consequences of changes of political and/or economic systems in the countries or regions in question. Examples are the Economies in Transition in Eastern Europe and South Africa.

| Rank | Country | 1995/96 | | 1985/86 | |
|---|---|---|---|---|---|
| | | Papers | Share | Papers | Share |
| 1 | Thailand | 1131 | 64.2% | 583 | 46.5% |
| 2 | Hungary | 5213 | 50.3% | 4670 | 26.5% |
| 3 | Portugal | 2870 | 50.1% | 813 | 37.8% |
| 4 | Czech Republic | 5587 | 49.1% | n. a. | [18.9%][1] |
| 5 | Switzerland | 20872 | 47.5% | 13506 | 32.1% |
| 6 | Poland | 12374 | 45.7% | 9261 | 20.2% |
| 7 | Chile | 2496 | 45.4% | 1557 | 25.5% |
| 8 | Belgium | 14695 | 45.0% | 9009 | 28.1% |
| 9 | Venezuela | 1137 | 44.9% | 733 | 30.8% |
| 10 | Romania | 2069 | 44.7% | 1301 | 15.0% |
| 11 | Slovenia | 1264 | 44.6% | n. a. | n. a. |
| 12 | Slovakia | 2815 | 44.3% | n. a. | [18.9%][1] |
| 13 | Denmark | 11809 | 43.3% | 8387 | 24.2% |
| 14 | Croatia | 1401 | 43.0% | n. a. | n. a. |
| 15 | Mexico | 4960 | 42.6% | 1997 | 29.9% |
| 16 | Austria | 9479 | 42.6% | 5439 | 23.8% |
| 17 | Brazil | 9417 | 41.7% | 3918 | 26.9% |
| 18 | Bulgaria | 2503 | 40.4% | 2611 | 20.9% |
| 19 | Ireland | 3162 | 40.3% | 1807 | 25.0% |
| 20 | Norway | 7131 | 40.0% | 5129 | 23.4% |
| 21 | Sweden | 23698 | 39.0% | 17143 | 21.9% |
| 22 | Greece | 5556 | 37.5% | 2629 | 25.4% |
| 23 | Hong Kong | 4191 | 37.5% | 1179 | 23.0% |
| 24 | Israel | 14067 | 37.1% | 11142 | 25.0% |
| 25 | Finland | 10361 | 35.6% | 6143 | 19.3% |
| 26 | Netherlands | 29773 | 35.4% | 18153 | 19.8% |
| 27 | France | 73925 | 34.2% | 47640 | 20.3% |
| 28 | Belarus | 1653 | 33.6% | n. a. | n. a. |
| 29 | Germany | 93683 | 33.3% | [58164][3] | [19.4%][3] |
| 30 | Italy | 46757 | 33.1% | 23913 | 21.1% |
| 31 | Argentina | 5167 | 32.0% | 3108 | 13.2% |
| 32 | Ukraine | 6691 | 31.3% | n. a. | n. a. |
| 33 | New Zealand | 5967 | 31.3% | 4729 | 15.8% |
| 34 | Egypt | 3266 | 31.0% | 2409 | 21.9% |
| 35 | Yugoslavia | 1326 | 30.8% | 2387 | 30.1% |
| 36 | Canada | 54369 | 30.6% | 43001 | 18.6% |
| 37 | Spain | 29538 | 30.0% | 10409 | 15.1% |
| 38 | PR China | 18861 | 28.8% | 6442 | 23.2% |
| 39 | Singapore | 2676 | 28.8% | 760 | 23.7% |
| 40 | UK | 110898 | 27.2% | 86721 | 14.4% |
| 41 | South Africa | 5448 | 27.0% | 5893 | 11.8% |
| 42 | South Korea | 10007 | 26.8% | 1221 | 27.3% |
| 43 | Australia | 30139 | 26.4% | 21200 | 14.5% |
| 44 | Russia | 44664 | 25.5% | n. a. | [3.3%][2] |
| 45 | Saudi Arabia | 1797 | 23.7% | 1173 | 26.5% |
| 46 | Turkey | 4798 | 21.4% | 926 | 25.9% |
| 47 | USA | 403056 | 18.1% | 340275 | 9.5% |
| 48 | Taiwan | 11594 | 17.5% | 1883 | 23.5% |
| 49 | India | 21449 | 15.2% | 21335 | 8.5% |
| 50 | Japan | 108019 | 14.4% | 67234 | 7.3% |

[1] Czechoslovakia [2] Soviet Union [3] without GDR

*Figure 5.4 Change of national publication output and share of international co-publications (all fields combined, 1985/86 vs. 1995/96)*

Frequently used measures for the strength of links are *Salton's* measure and the *Jaccard Index*. In addition, the extent of multi-nationality of collaboration can be studied. The

mathematical-methodological background of these measures will be introduced in the following section. Nevertheless, two examples for the use of *Salton's* measure in the context of international co-publication structures will be given at this place. Figures 5.5 and 5.6 visualise networks of international co-publications on the basis of the strength of collaboration links between country pairs. Multinational links have been split into the corresponding set of bilateral links. One has, consequently, to distinguish between the number of *co-publications* and of *co-authorship links*.

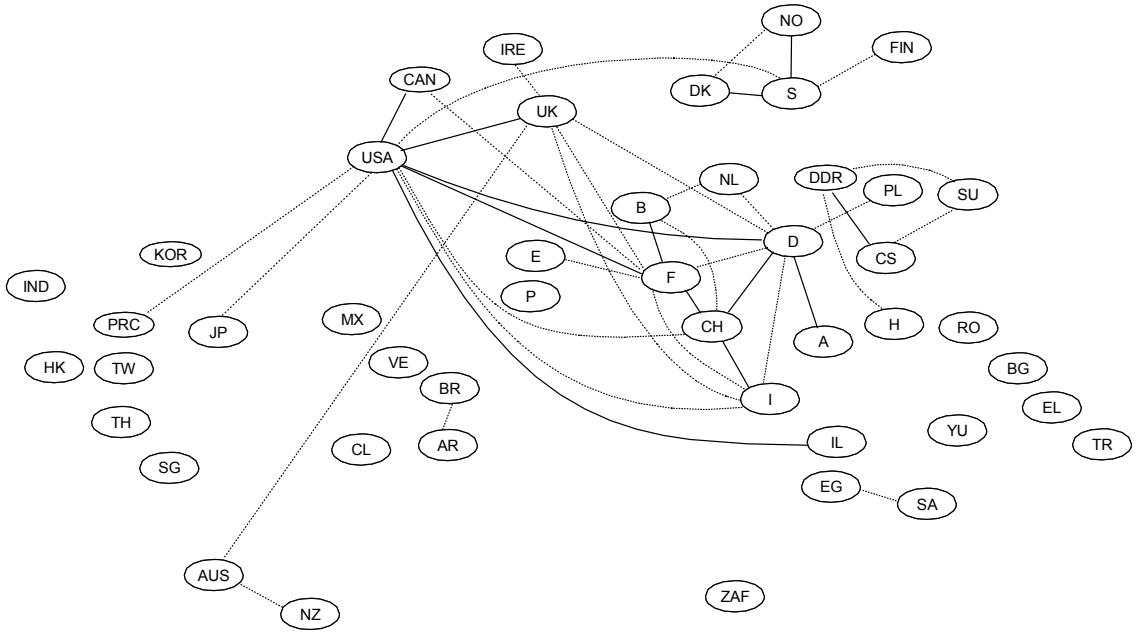



*Figure 5.5 Co-authorship map for 50 most active countries in all fields combined in 1985/86 based on Salton's measure (dotted line ≥ 1.5%, solid line ≥ 2.5%)*

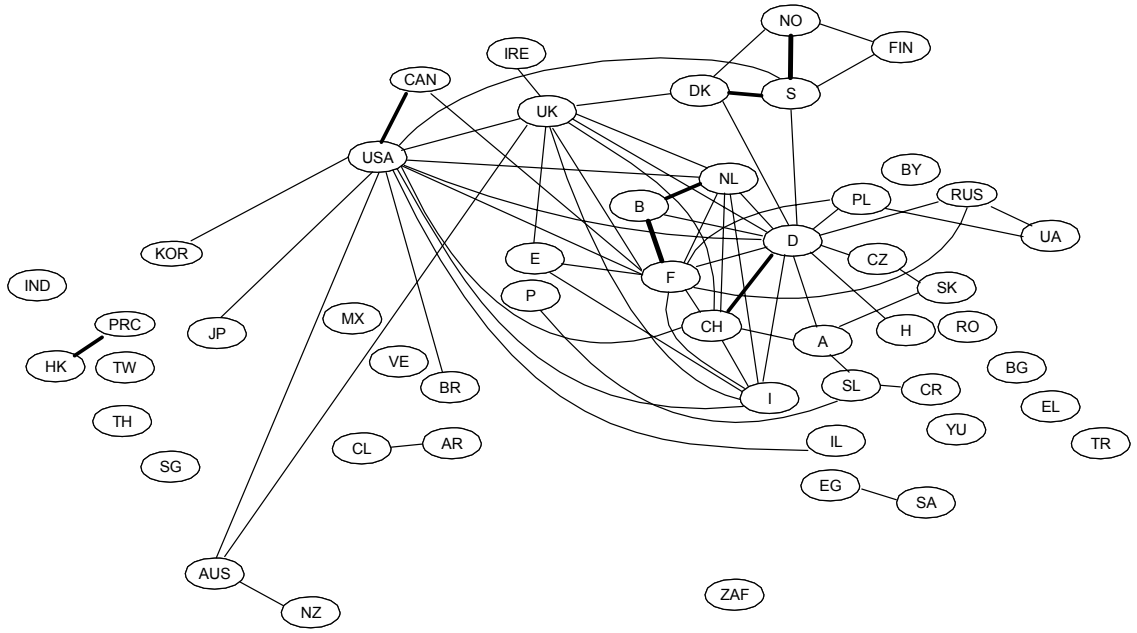



*Figure 5.6 Co-authorship map for 50 most active countries in all fields combined in 1995/96 based on Salton's measure (solid line ≥ 2.5%, thick line ≥ 5%)*

The two Figures 5.5 and 5.6 redrawn from *Glänzel* (2001) show that the share of international papers has grown in most countries; it also substantiates that international collaboration has intensified and the density of the network has increased. Moreover, the structure itself has changed. Although the two examples are almost self-explaining, some comments might help their interpretation.

- International scientific collaboration in the first period was not as intense as ten years later, therefore, lower thresholds ($r_{ij}$) of Salton's measure had to be used in Figure 5.5. As the lower threshold for medium (very) strong links, $r_{ij} = 1.5\%$ (2.5%) has been chosen. Values of Salton's measure exceeding 5% did not occur in 1985/86. The position of the countries on the map is intended to reflect the 'natural geographic order' as much as possible, and to express, at the same time, the structure defined by the co-authorship links. Six clusters of unequal size, namely, a big one including Western Europe, USA and Canada and two smaller ones with the Scandinavian and the Eastern European countries, respectively, can be found. There are further clusters, one including Australia and New Zealand, one consisting of Egypt and Saudi Arabia and a sixth one with Brazil and Argentina.
- The co-publication map presented in Figure 5.6 shows a quite different situation ten years later. Since the intensity of links has increased considerably from the 80s to the 90s, the corresponding thresholds had to be modified. The dense network of links had otherwise made the map unintelligible. Links with strength below 2.5% have, therefore, not been plotted. In addition, co-publication links stronger than 5% are represented by thick lines (see Figure 5.6). The changes are striking. First, the overall strength of links has increased; thus dotted lines in Figure 5.5 are regularly replaced by solid ones in Figure 5.6. Second, the network of co-publication links became denser although the lower threshold of 1.5% has been omitted in 1995/96. Third, a structural change can be observed. The Arabian cluster is still isolated, and has not changed. The South American cluster has undergone some structural changes. A tiny new cluster formed by the P.R. China and Hong Kong arose in Far East. The link between these two countries (5.9%) is one of the strongest in the period under study (a couple of years before the crown colony returned to China). The biggest cluster includes Europe, the USA and Canada. A strongly cross-linked EU cluster connected to the USA, a coherent Scandinavian cluster connected through Denmark and Sweden with the rest of the North American/European main cluster and a loosely connected Central/Eastern-European cluster joined to the main cluster through Germany and – as a new development – Poland playing jointly with Russia the role of a newly-fledged node in Eastern Europe.

Co-authorship is a symmetric (bi- or multidirectional) phenomenon; therefore, most bibliometric measures of international collaboration derived from co-publication analysis are symmetric. The maps presented in Figures 5.5 and 5.6 reflect mutual links but do not reveal anything about specific unidirectional 'affinities' of a country for co-authorship with other countries. However, it is possible to create 'asymmetric' measures of 'co-authorship affinity' to characterise the relative 'importance' of other countries for countries under study. With the help of the shares of joint papers with other countries in all co-publications and their share in the world total, a certain asymmetry can be uncovered in case of mutual affinity. Figure 5.7 presents the bi-literal codes of the 10 most important partner countries of Germany, France, USA and Japan (column *A*), the number of joint papers in all fields combined (column *B*), the percentage share of joint papers in the internationally co-authored papers of the selected

country (column *C*) together with the percentage share of the total number of publications of the same set of countries in the world total minus the number of publications of the selected country, Germany, France, USA and Japan, respectively (column *D*). The latter two values are identical if a country is exactly as important for the country under study as it is for the rest of the world. The values in column *D* may, in a way, serve as expectation for the *C* values. The deviation from this ideally balanced situation is in reality often considerable. There is, for instance, a clear asymmetry in the collaboration link Germany–Japan, since Germany is more important for Japan as expected, but Japan's importance for Germany is significantly below expectation. The same applies to the link USA–Japan. There is also a slight asymmetry in the link between USA and Germany, whereas the link between Germany and France can be characterised as a well-balanced mutual relationship. The examples presented in Figure 5.7 are taken from the study by *Glänzel* (2001) on national characteristics in international scientific co-authorship relations.

| Germany | | | | France | | | | USA | | | | Japan | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | **B** | **C** | **D** | **A** | **B** | **C** | **D** | **A** | **B** | **C** | **D** | **A** | **B** | **C** | **D** |
| **US** | 9381 | 30.03% | 37.72% | **US** | 6701 | 26.47% | 37.04% | **DE** | 9381 | 12.84% | 12.34% | **US** | 7268 | 46.67% | 38.23% |
| **UK** | 3780 | 12.10% | 10.38% | **DE** | 3528 | 13.94% | 8.61% | **UK** | 9296 | 12.73% | 14.61% | **DE** | 1499 | 9.63% | 8.89% |
| **FR** | 3528 | 11.29% | 8.77% | **UK** | 3295 | 13.02% | 10.19% | **CA** | 8703 | 11.91% | 7.16% | **UK** | 1479 | 9.50% | 10.52% |
| **RU** | 2626 | 8.41% | 4.18% | **IT** | 2547 | 10.06% | 4.30% | **JP** | 7268 | 9.95% | 14.23% | **CA** | 999 | 6.41% | 5.16% |
| **CH** | 2567 | 8.22% | 1.95% | **CH** | 1820 | 7.19% | 1.92% | **FR** | 6701 | 9.17% | 9.74% | **FR** | 885 | 5.68% | 7.01% |
| **IT** | 2242 | 7.18% | 4.38% | **ES** | 1785 | 7.05% | 2.71% | **IT** | 5112 | 7.00% | 6.16% | **CN** | 841 | 5.40% | 1.79% |
| **NL** | 1946 | 6.23% | 2.79% | **BE** | 1635 | 6.46% | 1.35% | **CH** | 3145 | 4.31% | 2.75% | **KR** | 619 | 3.97% | 0.95% |
| **JP** | 1499 | 4.80% | 10.11% | **CA** | 1593 | 6.29% | 5.00% | **NL** | 3121 | 4.27% | 3.92% | **IT** | 606 | 3.89% | 4.44% |
| **AT** | 1468 | 4.70% | 0.89% | **RU** | 1445 | 5.71% | 4.10% | **IL** | 2976 | 4.07% | 1.85% | **RU** | 581 | 3.73% | 4.24% |
| **SE** | 1259 | 4.03% | 2.22% | **NL** | 1326 | 5.24% | 2.74% | **AU** | 2859 | 3.91% | 3.97% | **AU** | 518 | 3.33% | 2.86% |

*Figure 5.7 Co-authorship affinity for four selected countries ranked by share of joint papers*
*(A = Country, B = number of joint papers, C = Share of joint papers in all international papers, D = Share of partner country in the world total minus country under study)*

Figure 5.8, finally, gives an outline of a possible organisation and the sketch of a methodological scheme of a complex analysis of international co-publication patterns according to the above-mentioned study by *Glänzel* (2001).

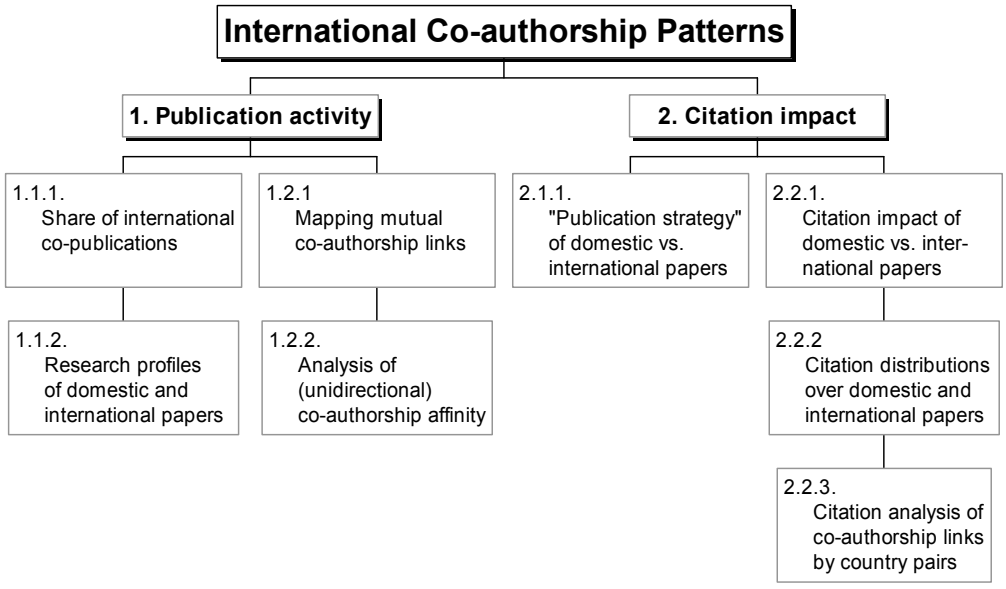



*Figure 5.8 Sketch of a possible methodological scheme of the analysis of international co-publication patterns*

# 6. INDICATORS AND ADVANCED DATA-ANALYTICAL METHODS

The previous section has already shown that bibliometric analyses might go far beyond direct comparison of indicator values and 'linear ranking'. The example of co-publication networks and unidirectional collaboration links, as well as the distributional approach to 'scientific productivity' and citation impact have shown that two- and multi-dimensional presentation might become necessary to describe underlying phenomena in an appropriate way. This requires advanced data-analytical techniques. Since statistical standard methods as provided by statistical programme packages can often be used (for instance, cluster analysis and multidimensional scaling), we will restrict this section to the underlying bibliometric rudiments yielding measures that might serve as input of statistical standard techniques. Statistical groundwork of such analyses is multivariate data-analysis techniques, that is, methods allowing the simultaneous analysis of quantitative relations between several variables. The aim of these analyses is, in particular, grouping (clustering) of bibliometric elements or units on the basis of similarity properties and/or measuring the 'distance' between them. Most bibliometric measures serving as a basis of these data-analytical methods can be derived from *bibliometric transaction matrices*. The visualisation of the results is called *bibliometric mapping.* Dynamic maps are used to monitor the change of structures in time.

## *6.1 Bibliometric transaction matrices*

Transaction matrices are square matrices of statistics giving transactions between bibliometric elements (e.g. papers or patents) and units (such as journals, countries etc.). These matrices have first been analysed by *Price* in 1981.
Joint references and citations, co-authorship, citation processes, attendance at conferences can be considered bibliometric transactions. The effect of undefined or dominant self-transaction is the main problem in the analysis of bibliometric square matrices, and has lead to the development of special statistical techniques. Figure 6.1 presents a fictitious example for symmetric bibliometric transaction matrix. The number of references shared by two papers, the number of joint articles of two units or the number of co-citations express symmetric relations because the relation between two elements or units is not directed in such cases.

| $n \times n$ | $p_1$ | $p_2$ | $p_3$ | … | $p_j$ | … | $p_n$ |
|---|---|---|---|---|---|---|---|
| $p_1$ | 17 | 0 | 0 | … | 2 | … | 1 |
| $p_2$ | 0 | 9 | 3 | … | 1 | … | 0 |
| $p_3$ | 0 | 3 | 56 | … | 0 | … | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| $p_j$ | 2 | 1 | 0 | … | 14 | … | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| $p_n$ | 1 | 0 | 1 | … | 0 | … | 22 |

*Figure 6.1 Example for a symmetric transaction matrix with dominant main diagonal (Matrix representing the number of joint references in a given paper set)*

The information flow expressed by citations given by one unit to another one is not necessarily reflecting mutual relationship. The same applies to the attendance of countries at international conferences. Cross-citations matrices and attendance at meetings or mobility are typical examples for non-symmetric transaction matrices. An example taken from *Schubert* et al. (1983) is shown by Figure 6.2.

| *n*×*n* | US | UK | DE | … | SU | … | IN |
|---|---|---|---|---|---|---|---|
| US | 4026 | 310 | 350 | … | 80 | … | 56 |
| UK | 617 | 988 | 220 | … | 48 | … | 32 |
| DE | 385 | 152 | 436 | … | 44 | … | 9 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| SU | 22 | 8 | 17 | … | 113 | … | 3 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| IN | 159 | 21 | 28 | … | 6 | … | 242 |

*Figure 6.2 Example for a non-symmetric transaction matrix with dominant main diagonal (Matrix representing the number of attendees at international conferences)*

Figure 6.3, finally presents an example for incomplete transaction matrices. The number of internationally co-authored publications between a country and itself country is not defined. Nevertheless, the missing main diagonal of the transaction matrix is often replaced by the total number of papers of the corresponding country in order to be able to calculate Salton's measure of the Jaccard Index on the basis of the matrix. All elements of the main diagonal of the resulting 'similarity' matrix are equal to 1, and have no particular meaning and interpretation.

| *n*×*n* | US | UK | DE | … | RU | … | IN |
|---|---|---|---|---|---|---|---|
| US | • | 328 | 374 | … | 128 | … | 78 |
| UK | 328 | • | 74 | … | 44 | … | 5 |
| DE | 374 | 74 | • | … | 95 | … | 22 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| RU | 128 | 44 | 95 | … | • | … | 3 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| IN | 78 | 5 | 22 | … | 3 | … | • |

*Figure 6.3 Example for a symmetric transaction matrix with undefined main diagonal (Matrix representing the number of international co-publications)*

As mentioned above, bibliometric transaction matrices can be symmetric or non-symmetric. The latter ones can be symmetrised by the following transformation: $\mathbf{B} = ½\cdot(\mathbf{A} + \mathbf{A}^T)$, where **A** is the original non-symmetric matrix and **B** is the symmetrised transaction matrix.

## *6.2 Bibliographic coupling and co-citation analysis*

The most important bibliometric method based on reference literature is *bibliographic coupling*. The concept of *bibliographic coupling* was proposed more than three decades ago by *Kessler* (1964). In a comprehensive validation study by *Vladutz* and *Cook* (1984) could be shown

> *"that the utilization of bibliographical coupling in very large citation databases is practically feasible and that in 76% of the cases it yields valid results by providing for each input publication 2-3 other publications from the same year, that are closely related by subject".*

Therefore, the method of bibliographic coupling may prove a valuable alternative to other mapping methods, especially, because this technique is immediately available and applicable after publication of a body of literature that does not only contain cited documents.

*Bibliographic coupling* is often confused or wrongly considered equivalent with *co-citation* analyses. The latter ones analyse reference pairs, that is, cited papers, whereas bibliographic-coupling methods proceed from those citing papers that share items in their reference lists. Both bibliographically linked publications and co-citation links are assumed to form clusters representing the same or at least related research topics.

We introduce the two techniques with the help of a model suggested by *S.K. Sen* and *S.K. Gan* in 1983 and generalised by *Glänzel* and *Czerwon* in 1996. The model can briefly be described as follows.
The total set of scientific literature generates a Boolean vector space which is, particularly, defined by the relationship of papers published in a given period and the set of all references cited by them. Assume that all $m$ references cited in all $n$ papers in question are arranged and indexed in some order. The elements of the space are Boolean vectors representing the publications, their $j$-th component ($1 \le j \le m$) takes then the values 1 or 0, according as the paper cites the references $j$ or not. The document–reference assignment of the total paper set published in a given period and indexed in some sequence can then be represented by a huge $n \times m$ Boolean matrix $\underline{\mathbf{A}}$. Without the loss of generality we assume that the rows consist of the publication (document) vectors $\mathbf{d}_i$ ($1 \le i \le n$) and the columns represent the reference vectors $\mathbf{r}_j$ ($1 \le j \le m$). Figure 6.3 presents an example matrix.
The matrix $\underline{\mathbf{A}}$ in its present form does, so far, not reveal anything about bibliographic links and their strength. *Sen* and *Gan* have, therefore, suggested using the *Coupling Angle* (*CA*) as a coupling measure. It is defined as the cosine of two Boolean vectors $\mathbf{d}_i$ and $\mathbf{d}_j$ that can be obtained from their scalar product as follows

$$CA(\mathbf{d}_i \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i| \cdot |\mathbf{d}_j|} \ .$$

The Coupling Angle *CA* takes then the value 1 if two Boolean vectors are parallel, and $CA = 0$ if they are rectangular. The strength of bibliographic coupling of two papers can thus be visualised as the angle between the corresponding Boolean vectors. This may help to find appropriate thresholds based on its geometrical interpretation. Papers represented by rectangular vectors are independent; those represented by parallel vectors are concerned with identical topics. Two documents may then be considered to be concerned with a related topic

if the angle between the vectors representing the documents does not exceed a given angle φ ($0° \le \varphi < 90°$). For example, the choice of $\varphi = 60°$ results in a threshold $CA = 0.5$, and the angle $\varphi = 75°$ corresponds to a threshold $CA \approx 0.25$.

| | | Cited papers (references) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n \times m$ | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ | $r_6$ | ⋯ | $r_j$ | ⋯ | $r_m$ |
| Citing papers | $d_1$ | 1 | 0 | 0 | 1 | 0 | 0 | ⋯ | 0 | ⋯ | 1 |
| | $d_2$ | 0 | 0 | 0 | 0 | 1 | 0 | ⋯ | 1 | ⋯ | 0 |
| | $d_3$ | 0 | 1 | 1 | 0 | 0 | 0 | ⋯ | 0 | ⋯ | 1 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| | $d_i$ | 1 | 0 | 1 | 1 | 0 | 1 | ⋯ | 1 | ⋯ | 0 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ |
| | $d_n$ | 1 | 1 | 0 | 0 | 0 | 1 | ⋯ | 0 | ⋯ | 0 |

*Figure 6.4 A hypothetical Boolean matrix representing the relationship publications–references*

*Glänzel* and *Czerwon* (1996) have shown that in the above case the Coupling Angle is identical with *Salton's measure* ($r_{ij}$). Thus we have $CA(\mathbf{d_i d_j}) = r_{ij}$ and a 'geometric' interpretation of *Salton's measure* in terms of the angles between two Boolean vectors. Matrices representing bibliographic links and their strength can be readily derived from the Boolean matrix $\underline{\mathbf{A}}$ as introduced above. The product matrix $\underline{\mathbf{B}} = \underline{\mathbf{A}} \cdot \underline{\mathbf{A}}^{\mathrm{T}}$ provides the number of joint references of the given papers. An example for such a 'bibliographic-link' matrix is given in Figure 6.4. Finally, the matrix $\underline{\mathbf{R}} = \mathrm{Diag}(\underline{\mathbf{B}})^{-½} \cdot \underline{\mathbf{B}} \cdot \mathrm{Diag}(\underline{\mathbf{B}})^{-½} = [r_{ij}] = [CA(\mathbf{d_i d_j})]$ contains *Salton's measure* of the strength of the bibliographic link of all pairs of papers.
A more popular definition regards *Salton's measure* as the ratio of the number of joint references and the geometric mean of the number of references of the two papers concerned. Bibliographic coupling is without any restriction a symmetrical relation. Moreover $r_{ij}$ takes the value 1 for the strength of the bibliographic link of any paper with itself. In contrast to several other bibliometric phenomena such as, e.g., international collaboration, this makes sense because the relationship of any paper to itself must be considered maximum. These properties guarantee that Salton's measure is an appropriate measure of the strength of bibliographic coupling.

The method of *co-citation clustering* was introduced independently by *Small* (1973) and *Marshakova* (1973). The matrix the elements of which present co-citation links can be derived from the same Boolean matrix $\underline{\mathbf{A}}$. Here the *Coupling Angle* is defined as the cosine of two Boolean column vectors $\mathbf{r}_i$ and $\mathbf{r}_j$, i.e.,

$$CA(\mathbf{r_i}\mathbf{r_j}) = \frac{\mathbf{r_i} \cdot \mathbf{r_j}}{|\mathbf{r_i}| \cdot |\mathbf{r_j}|} \ .$$

The product matrix $\underline{\mathbf{C}} = \underline{\mathbf{A}}^{\mathrm{T}} \cdot \underline{\mathbf{A}}$ then provides the number of co-citations of the given papers. The matrix $\underline{\mathbf{S}} = \mathrm{Diag}(\underline{\mathbf{C}})^{-½} \cdot \underline{\mathbf{C}} \cdot \mathrm{Diag}(\underline{\mathbf{C}})^{-½} = [r_{ij}] = [CA(\mathbf{r_i}\mathbf{r_j})]$ contains *Salton's measure* of the strength of the co-citation link of all pairs of papers. Note, that $\underline{\mathbf{C}}$ is not simply the transpose of the corresponding bibliographic coupling matrix $\underline{\mathbf{B}}$, i.e., $\underline{\mathbf{C}} \neq \underline{\mathbf{B}}$.

An alternative measure for the strength of links is the *Jaccard Index.* It was originally defined as a measure of the relative intersection of finite sets. In particular, we have

$$J_{ij} = \frac{\mathbf{A}_i \cap \mathbf{A}_j}{\mathbf{A}_i \cup \mathbf{A}_j} = \frac{n_{ij}}{n_i + n_j - n_{ij}} ,$$

where $\mathbf{A}_i$ and $\mathbf{A}_j$ are finite sets with cardinality $n_i$ and $n_j$, respectively, and the cardinality of their intersection is denoted by $n_{ij}$. Also $J_{ij}$ takes its values in the interval [0, 1], where $J_{ij} = 0$ means that the two sets are disjoint, and $J_{ij} = 1$ means that the two sets are identical. In the case of bibliographic coupling each set $\mathbf{A}_i$ and $\mathbf{A}_j$ represents the set of items in a reference list.

Up to now the technique of bibliographic coupling has been applied very rarely for purposes of research evaluation. This fact is surprising because the bibliographic coupling concept has obviously advantages compared with co-citation clustering.

- The most important one is that just published papers that are closely related by bibliographic coupling links can provide snapshots of early stages of a speciality's evolution. By contrast, it may take time before a 'critical mass' of papers on a new research topic is created that is needed to produce the highly cited publications on which the co-citation mapping is based (*Hicks*, 1987).
- From the viewpoint of information retrieval, *co-citation clustering* results in the restriction to frequently cited papers, whereas *bibliographic coupling* extends to practically all publications.

*Sharabchiev* (1988) has shown that both techniques provide comparable results of the structure of research front specialities and complement each other. ISI is using combinations of the two worlds for their products, for instance, *SciViz* (cf. *Small*, 1998).

*Glänzel* and *Czerwon* (1996) have used bibliographic coupling also to identify *core documents* that represent 'hot' and research-front topics. This method makes it possible to apply bibliographic-coupling techniques to both research evaluation and information retrieval.

### *6.3 Co-word, Co-heading and Co-author Clustering Techniques*

These techniques are based on the analysis of co-occurrences of terms, keywords or subject headings. Most prominent large-scale product for co-word analysis is LEXIMAP and its derivates; the used algorithm is LEXINET. This method has been developed in France by *Turner, Callon* and *Courtial* (see, for example, *Callon* et al., 1983, 1986, *Turner* et al., 1988, *Law* and *Whittaker*, 1992 and *Courtial*, 1994). Co-word clustering is also a standard technique at ISI. Co-word analysis is based on frequency analyses of co-occurrence of keywords

extracted from titles, abstracts or text, in general. Since its fields of application are manifold it has already evolved to an own discipline.

*Todorov* (1990) has developed a co-heading analysis that is based on co-occurrence of subject headings used in specified bibliographic databases. He has used the subject classification of the INSPEC database. His methodology has certain database-specific limitations.

Unlike co-word, co-citation and bibliographic coupling techniques that are above all designed to describe the structure of science and it evolution at the macro and meso level, co-author clustering and author co-citation analysis (ACA) are interesting bibliometric methods to reveal structures at the micro and meso levels (*White* and *McCain*, 1998). However, it should be mentioned that ACA has a serious shortcoming. It is based on first-author co-citation. Both the sequence of co-authors and the strength of collaboration links are somewhat distorting the results.

Several authors have shown that the combination of several different methods essentially improves efficiency and validity of results (cf., *Sharabchiev*, 1988, *Braam* et al., 1991).

### *6.4 Techniques of Matrix Analysis*

Basically, the following three techniques can be applied to the analysis of bibliometric transition matrices.

- Deviation of observations from expectation
- Calculation of distance and similarity measures
- Decomposition of the matrix

The derived measures can be used as bibliometric indicators themselves (e.g., *Salton's* measure, *Jaccard* Index), or serve as a basis of multivariate analyses such as cluster analysis, multidimensional scaling, quasi-correspondence analysis etc. These techniques, which are preferably applied to mapping of the structure and dynamics of science, can be considered an extension of bibliometric indicators.

Scientometrics transaction matrices have first been analysed by *Price* (1981). *Price* has shown that distorted or censored main diagonals of matrices can uniquely be reconstructed. Let $\underline{\mathbf{A}} = (a_{ij})$ be an $n \times n$ square matrix with $n > 2$. The basis assumption is that the transaction matrix $\underline{\mathbf{A}}$ represents an *independence model*, that is, the transaction levels between the entities or units are determined by their size only. Consequently, we can write $\underline{\mathbf{A}} = \underline{\mathbf{p}} \cdot \underline{\mathbf{q}}^{\mathrm{T}}$ for some appropriate vectors $\underline{\mathbf{p}}$ and $\underline{\mathbf{q}}$. In particular, we have $a_{ij} = p_i \cdot q_j$ for all indices $I$ and $j$ ($1 \le i, j \le n$). In case of dominant main diagonal, this model fails, and the $a_{ii}$ values might thus be considered distorted, and can be omitted. Thus we assume that the elements of the main diagonal are unknown. The products of the *known* elements over rows and columns are denoted by $a_{i\times}$' and $a_{\times j}$', respectively. $a_{\times\times}$' denotes the grand total product of the matrix omitting the unknown diagonal. Since the vectors $\underline{\mathbf{p}}$ and $\underline{\mathbf{q}}$ and thus their components $p_i$ and $q_j$ are unknown, the diagonal elements of $\underline{\mathbf{A}}$ cannot be immediately reconstructed. However, since under the above assumption $a_{\times\times}' = \{\Pi_i \; p_i \cdot \Pi_j \; q_j\}^{n-1}$ and for the corresponding sub-matrices $\underline{\mathbf{A}}_{kk}$, $a_{\times\times}'_{kk} = \{\Pi_{i \neq k} \, p_i \cdot \Pi_{j \neq k} \, q_j\}^{n-2}$, we have

$$p_k \cdot q_k = (a_{\times\times}')^{1/(n-1)}/(a_{\times\times}'_{kk})^{1/(n-2)} \text{ for all } k = 1,2, \ldots, n.$$

This formula can be a bit simplified by using the following substitution: $a_{\times\times}'_{kk} = a_{\times\times}'/(a_{k\times}' \cdot a_{\times k}')$. Thus, we finally have the following solution

$$p_k \cdot q_k = (a_{k\times}' \cdot a_{\times k}')^{1/(n-2)}/(a_{\times\times}')^{1/((n-1)\cdot(n-2))} \; ; \; k = 1,2, \ldots, n.$$

Hence we obtain the reconstructed matrix $\underline{\mathbf{A}}^* = (a_{ij}^*)$ with $a_{ij}^* = a_{ij}$ for $i \neq j$ and $a_{ii}^* = (a_{i\times}' \cdot a_{\times i}')^{1/(n-2)}/(a_{\times\times}')^{1/((n-1)\cdot(n-2))}$.
It is known that the maximum likelihood estimator of the expected number of transactions $e_{ij}$ can be obtained from the observed transactions $a_{ij}$ in the following manner.

$$e_{ij} = a_{i+}a_{+j}/a_{++} \, ,$$

where $a_{i+} = \Sigma_j \, a_{ij}$, $a_{+j} = \Sigma_i \, a_{ij}$ and $a_{++} = \Sigma_i \Sigma_j \, a_{ij}$. If the main diagonal is known and not dominant, there is from the statistical viewpoint no problem. Then the significance of the deviation of the observed transactions from the estimated ones can be evaluated on the basis of the following $\chi^2$-statistic

$$\chi^2 = \Sigma_i \Sigma_j \, (a_{ij} - e_{ij})^2/e_{ij}.$$

However, the task is not the evaluation of the goodness-of-fit, namely, of the significance of the deviation from the independence model but to measure the distance of individual observations from their expectations. This can be done with the help of *correspondence analysis* (CA). This method allows the joint plot of variables in the same diagram (cf. *Tijssen* et al., 1988). Figure 6.5 taken from *Tijssen* et al. (1988) presents a typical application of CA. Subject fields represented by PACS codes taken from the INSPEC database are jointly presented with journals in a CA diagram. This diagram largely reflects specialisation of scientific journals.

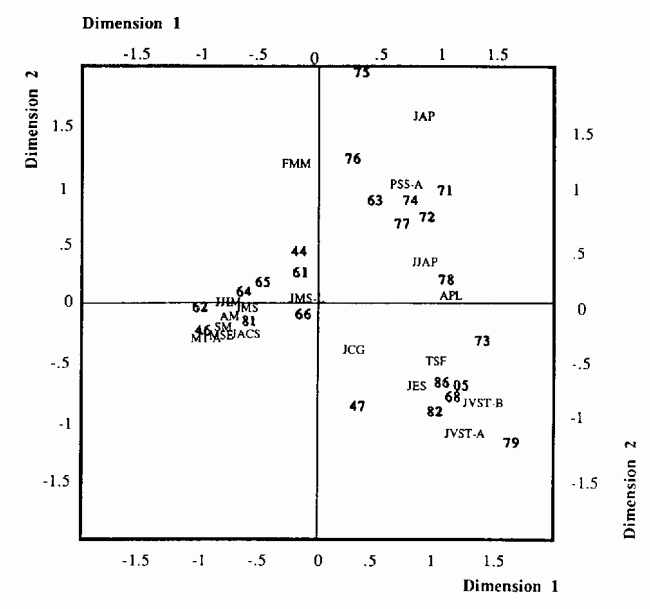



*Figure 6.5 Correspondence analysis of a journal-to-field matrix in physics in 1985 (according to Tijssen et al., 1988)*

The indipendence model cannot be assumed if the main diagonal of the transaction matrix is dominant. The method by *Price* (1981) does not allow the use of the Pearson statistic $\chi^2 = \Sigma_i \Sigma_j \ (a_{ij} - e_{ij})^2 / e_{ij}$ since the element of the main diagonal have been determined with the help of the off-diagonal elements. However, there are several other ways to solve this problem. The quasi-independence model offers one possibility. According to this model, elements of the matrix can be ignored and the independence model still fits on the remaining observations $a_{ij}$, where $i$ and $j$ are elements of a subset $L$ of the index set. Usually, $L = \{(i,j);\ i \neq j,\ i,j = 1,\ 2,\ \ldots\ ,\ n\}$ is assumed. A $\chi^2$-statistic can be constructed with restriction to $L$. The quasi-correspondence analysis (QCA) can then be conducted in analogous manner to the CA. A comprehensive description of the QCA can be found, for instance, in the papers by *de Leeuw* and *van der Heijden* (1985) and *Tijssen* et al. (1987). Figure 6.6 presents a typical application of QCA to cross-journal citation analysis. The main diagonal in cross-citation matrices is usually heavily distorted by the large share of self-citations.

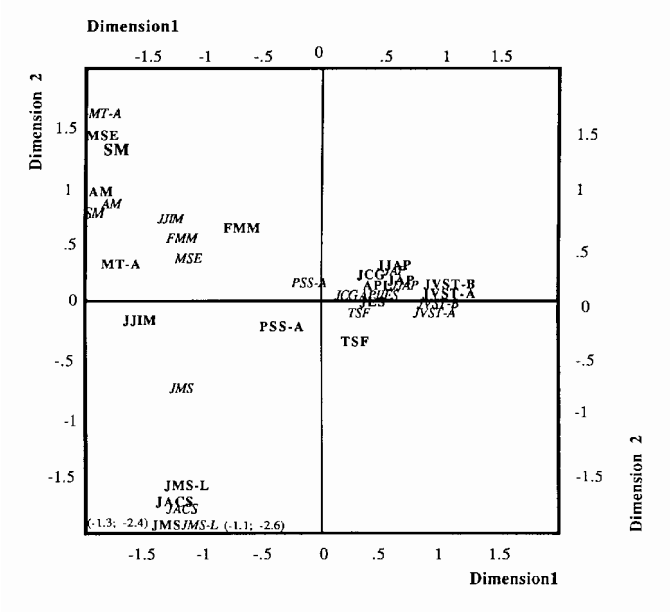



*Figure 6.6 Quasi-correspondence analysis of a cross-citation matrix in physics in 1985 (italics: citing journals, bold: cited journals; according to Tijssen et al., 1988)*

A second possibility to eliminate dominant main diagonals is an appropriate decomposition. In the paper on attendance at international scientific meetings *Schubert* et al. (1985) have suggested the following decomposition. The transaction matrix **A** can be decomposed into a sum of a dyadic product of appropriate vectors **p** and **q** representing the independence model and a diagonal matrix **C** representing surplus self-transactions. According to this approach, we have

$$\mathbf{A} = \mathbf{p} \cdot \mathbf{q}^T + \mathbf{C}.$$

Analogously to the previous solution, the algorithm is based on an iteration method. The results can be analysed on the basis of the deviation from their expectation, and be visualised in relational charts. Moreover, the elements of the main diagonal of the matrix **C** reveal interesting additional information on self-transactions.

# 7. THE BORDERLAND OF BIBLIOMETRIC RESEARCH

## *7.1. Linkage Between Science and Technology*

On Tuesday, 13 May, 1997 a spectacular article entitled “Study Finds Public Science is Pillar of Industry" appeared in the Science Times Section of the New York Times (*Broad*, 1997). This article was based on results published in a study for the National Science Foundation for which CHI traced more than 45,000 references from U.S. patents to the underlying research papers. *Narin* (CHI) found that more than 70 percent of the scientific papers cited on the front pages of U.S. Industry patents came from public science (science performed at universities, government labs, and other public agencies. This implies that public science has a direct, massive impact on industrial technology. *Narin* also found that this dependence is increasing rapidly, more than tripling in the six years from 1987/88 to 1993/94.

To measure the direct impact of science on technology with bibliometric means is a difficult task. Studies on the *citation gap* in applicable sciences (*Vanels* et al., 1989, *Jansz* and *LePair*, 1992, *Jansz*, 1999) have pointed to this problem. According to these studies communication patterns in these areas do not primarily rely on publications in scientific journals or, in some fields, in patents. The analysis of *patents* by bibliometric means can, however, help to study quantitative aspects of research at the borderline between science and technology. The combination between publication and patent analysis can be useful in such subject areas, and can thus improve the validity of the bibliometric studies.
Unlike the publication analysis where most methodological questions are solved, several methodological aspects of patent analyses are still controversy.
Patent analyses are based on relevant information that can be retrieved from the patent databases. The databases of the *United States Patent and Trademark Office* (USPTO) and of the *European Patent Office* (EPO) are probably the most frequently used databases for analyses of patent-literature links. Moreover, patent information can be also retrieved from the bibliographic database *Chemical Abstract* that indexes subject-relevant patents, too. In the following table, most important information used in bibliometrics is shown. This information is usually organised in corresponding search fields.

*Relevant information from patent databases:*

*Patents*

1. Patent identification
2. Names of inventors
3. Assignee
4. Addresses
5. References (patents and other publications)
6. Abstract
7. Classification

Figure 2.3 gives an example for patent information provided by the USPTO.

```
PATN
WKU  062637363
SRC  9
APN  4049543
APT  1
ART  286
APD  19990924
TTL  Electrostatically tunable resonance frequency beam utilizing a stress-sensitive film
ISD  20010724
NCL  27
ECL  1
EXP  Moller; Richard A.
NDR  7
NFG  16
INVT
NAM  Thundat; Thomas G.
CTY  Knoxville
STA  TN
INVT
NAM  Wachter; Eric A.
CTY  Oak Ridge
STA  TN
INVT
NAM  Davis; J. Kenneth
CTY  Kingston
STA  TN
ASSG
NAM  UT-Battelle, LLC
CTY  Oak Ridge
STA  TN
COD  02
CLAS
OCL   7351436
XCL   7351434
XCL   7351426
XCL  310309
XCL  361280
EDF  7
ICL  G01P 1500
FSC   73
FSS  514.36;514.34;514.26;514.17;514.18
FSC  310
FSS  309
FSC  361
FSS  280;283.1;287
UREF
PNO  5211051
ISD  19930500
NAM  Kaiser et al.
OCL   73  1D
UREF
PNO  5267471
ISD  19931200
NAM  Abraham et al.
OCL   73105
UREF
PNO  5719324
ISD  19980200
NAM  Thundat
UREF
PNO  5918263
ISD  19990600
NAM  Thundat
OREF
PAL  G. Y. Chen, et al "Adsorption-Induced Surface Stress & Its Effects on
--    Resonance Frequency of Microcaltilevers" J.Appl.Phys. 77 (8), Apr. 1995, 1-5.
PAL  M. Ilavsky et al, Responsive Gels: Volume Transitions I Editor: K. Dusek, 1993.
LREP
FR2  Marasco; Joseph A.
ABST
PAL  Methods and apparatus for detecting particular frequencies of acoustic
--    vibration utilize an electrostatically-tunable beam element having a
--    stress-sensitive coating and means for providing electrostatic force to
--    controllably deflect the beam element thereby changing its stiffness and
--    its resonance frequency. It is then determined from the response of the
--    electrostatically-tunable beam element to the acoustical vibration to
--    which the beam is exposed whether or not a particular frequency or
--    frequencies of acoustic vibration are detected.
GOVT
PAR  The United States Government has rights in this invention pursuant to
--    contract no. DE-AC05-96OR22464 between the United States Department of
--    Energy and Lockheed Martin Energy Research Corporation.
```

*Figure 7.1 Complete information about a patent by Thundat et al. (2001) according to the USPTO database*

Links between science and technology can basically studied on the basis of the following biblio-technometric relations.

- References to scientific publications in patents
- References to patents in scientific publications (so-called *reverse* patent citations)
- Authors-inventor relations

Links established through patent citations to scientific literature are considered most important, and is therefore most frequently applied (see, for instance, *Narin* et al., 1997, *Verbeek* et al., 2002). Nevertheless, also the reverse linkage, established by citation to patents in scientific literature uncovers interesting aspects of science-technology links (see, for instance, *Hicks*, 2000, *Glänzel* and *Meyer*, 2002). The high relevance of patents in chemistry research, among others, reflected by indexing patents in the *Chemical Abstracts* bibliographic database, could be also confirmed by *Glänzel* and *Meyer* (2002).
The third issue is probably the most promising one, however, it results in some, partially unresolved technical problems. The correct identification of inventors as authors of scientific publication in large-scale analyses is almost unfeasible. Nevertheless, at lower levels of aggregations or in specific domains, reliable studies of this relationship can be conducted (cf., *Noyons* et al., 1994).

### *7.2. New horizons: bibliometric methods in webometrics*

In bibliometrics publications are considered the elementary units of scientific information and the main source of indicators. The diversity of new patterns of communication on the electronic network and the heterogeneity of the internet, however, blurs the traditional frontiers between formal and informal communication. According to *Björneborn* and *Ingwersen* (2001) point out, that the web is an information space quite different from common scientific and professional databases, the similarities between electronic and print medium are often superficial. In particular, they write:

> *"Obviously, the breakthrough for everybody to express themselves, practically without control from authorities, to become visible world wide, also by linking to what pages one wants to link, to assume credibility by being 'there', and to obtain access to data, information, values and knowledge in many shapes of and degrees of truth, has generated an a reality of freedom of information, also in regions and countries otherwise poor of infrastructure."*

The most important difference between print media and the web is that time plays a different role on the web. An additional fundamental difference between print and web based analysis is the possibility of an almost continuous change of contents on the web (*Glänzel*, 2001/2003) resulting in a completely different notion of ageing of information.
Although the analogy between bibliometric and webometrics phenomena is limited, several bibliometric measures and models can be applied in webometrics, too. In the following, we briefly summarise possible web-based measures.

- The number of *visitors* of web sites can be described by point processes or pure birth processes being one of the rare cumulative process on the web.
- The frequency of *downloads* can be described by informetric models used in the context of copy requests.

- *Sitation* and *co-sitation* analyses have already been conducted. There are analogies but also deviations from the bibliometric (co-) citation models (*Rousseau*, 1997, *Björneborn* and *Ingwersen*, 2001, *Glänzel*, 2001/2003).
- *'Webographic coupling'* might be a promising tool for depicting structures of and identifying cluster in the web.
- The possible analysis of *(co-)sponsorship* has no real counterpart in bibliometrics.
- Studies of 'small world' phenomena (see, i.e., *Watts* and *Strogatz*, 1998) are more relevant in webometrics than in bibliometrics where little has been done in this area.
- Analysis and mapping of network structures on the web. In this context many formal statistical techniques also used in bibliometrics can be applied.

In all, we can conclude that structures on the Web are more complex than their bibliographic 'counterparts'. However, the main difference between print and web medium can be summarised as follows. Most bibliometric processes are cumulative since publications (except for the extremely rare cases of *retractions*) and citations are irreversible and bibliographic links cannot be removed if they have once been established. By contrast, the Web is in terms of both, *content* and *links* in permanent change.

# 8. INTRODUCTION INTO BIBLIOMETRIC TECHNOLOGY

## *8.1 Outlines of cleaning-up and computerised data processing of bibliographic data*

In the following we give an example for computerised data processing to visualise the way from bibliographic data to bibliometric indicators. To process all bibliometric indicators needed for the comparative studies, to calculate the necessary statistical functions, to fit the underlying frequency distributions and to estimate their parameters the development of appropriate bibliometric software systems is necessary. The first step in data processing is usually the process of downloading or extracting data from bibliographic databases. The downloaded data cannot immediately be processed into indicators. Due to spelling variances and errors, these data have to be carefully cleaned up. Basically four main sources of errors can be identified.

- the authors of the publications indexed in the database
- the editors of the journals covered by the database
- the database producer
- the user of the database

The authors themselves are responsible for many errors. Among these errors, we find, above all, misspelling, incomplete or wrong addresses and incorrect citations. These errors can practically not be corrected, except for the level of individual publications provided that the full text of all papers involved are available. The following example might illustrate this quite dramatic effect. The paper by *Schubert*, *Glänzel* and *Braun* (1989) "Scientometric datafiles. A Comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields, 1981-1985" published in *Scientometrics*, vol. **16,** pp. 3-478 has received 137 citations. Among those are 113 correct citations, whereas 24 citations were incorrect. The error caused by citing authors amounts to 17.5%. We can detect three particular types of errors; the most frequent was the incorrect or missing page number, followed by an incorrect publication year and the incorrect first author. The head of a research group or institute is sometimes erroneously assumed to be the first authors. All variances of the cited work that occurred in the Web of Science database are presented in Figure 8.1.

| Cites | 1st author | Journal | VOL | BP | PY |
|---|---|---|---|---|---|
| 113 | SCHUBERT A | SCIENTOMETRICS | 16 | 3 | 1989 |
| 3 | SCHUBERT A | SCIENTOMETRICS | 16 | 3 | *1988* |
| 1 | SCHUBERT A | SCIENTOMETRICS | 16 | 3 | *1987* |
| 2 | *BRAUN T* | SCIENTOMETRICS | 16 | 3 | 1989 |
| 12 | SCHUBERT A | SCIENTOMETRICS | 16 | *1* | 1989 |
| 1 | SCHUBERT A | SCIENTOMETRICS | 16 | *8* | 1989 |
| 1 | SCHUBERT A | SCIENTOMETRICS | 16 | *18* | 1989 |
| 1 | SCHUBERT A | SCIENTOMETRICS | 16 | *218* | 1989 |
| 1 | SCHUBERT A | SCIENTOMETRICS | 16 | *239* | 1989 |
| 1 | SCHUBERT A | SCIENTOMETRICS | 16 | *432* | 1989 |
| 1 | SCHUBERT A | SCIENTOMETRICS | 16 | | 1989 |

*Figure 8.1 Example for an incorrectly cited publication*

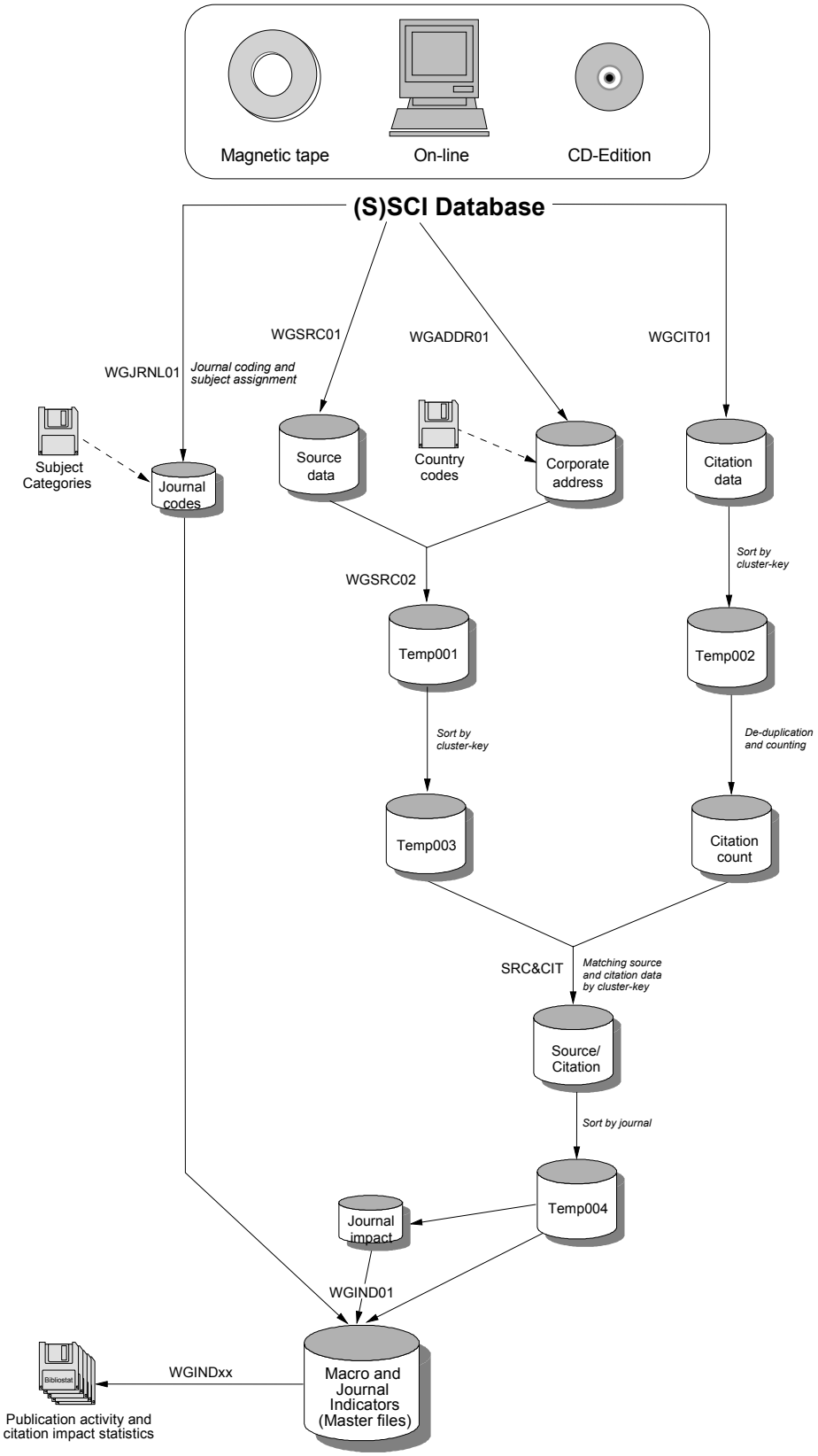



*Figure 8.2 Simplified flow chart of computer processing of bibliometric indicators*

Editorial policy is also responsible for a certain share of errors. Omitted addresses or truncated references are among such errors. These errors cannot be corrected by manual or computerised standard procedures.
Errors caused by the database producers are twofold; errors during data recording are hard to detect. However, these errors have from the statistical viewpoint not much weight. Systematic errors due to the policy in preparing the database have sometime sever consequences, but can often be corrected by users. The case of *Angewandte Chemie – International Edition* may just serve as one example (see, *Braun* and *Glänzel*, 1995, and *van Leeuwen* et al., 1997).
Unfortunately, there are also errors caused by the user. Standardisation of data processing techniques, the exchange of results among bibliometric research centres (cf. *Glänzel*, 1996) and co-operation as the best solution to developing common standards (cf. *Katz*, 1996) might help to reduce errors caused by users to the minimum.

Figure 8.2 and 8.3 present a overview of main steps of processing of bibliographic data to indicators.

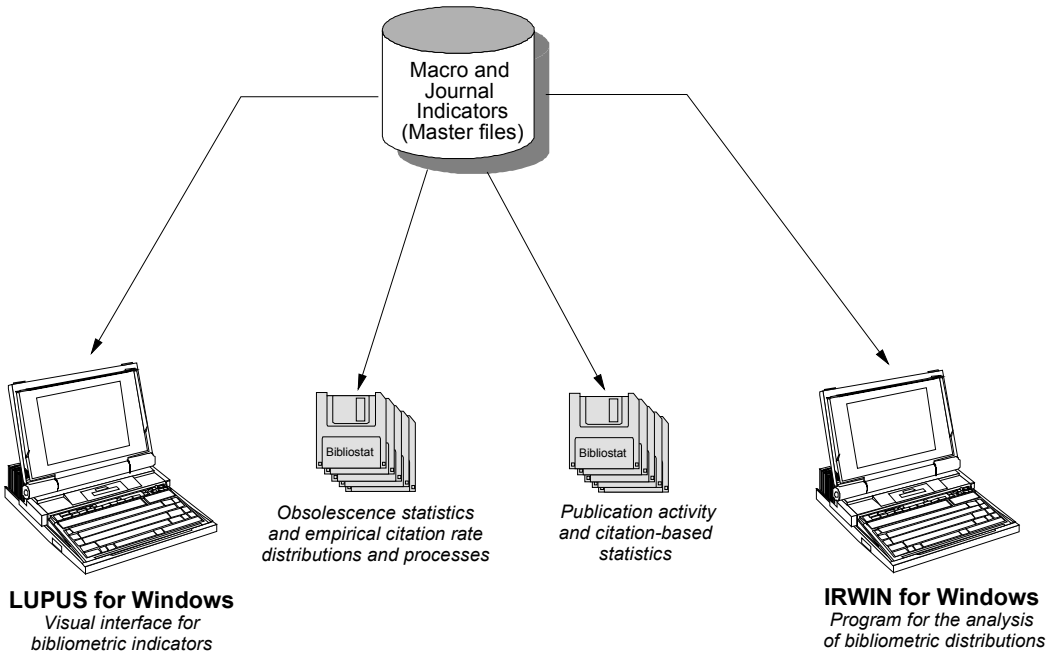



*Figure 8.3 Outline of statistical applications*

The program *Lupus – Visual Bibliometrics* developed by *Glänzel* in 1995 is designed for the use in addition to the CD-ROM edition of the (S)SCI of the Institute for Scientific Information (ISI). The software is compatible to the annual cumulations of the CD-Edition.
Visual Bibliometrics calculates and visualises basic and more advanced macro-indicators. Besides the main window five charts and maps can be chosen according as which indicator is to be visualised. The main window displays the most important bibliometric macro-indicators at the national level if the fields *Selected country*, *Selected field* and *Publication year* have valid entries. Besides the set of displayed bibliometric indicators a country profile summarises bibliometric performance characteristics. Figure 8.4 through 8.7 show the main features of this add-on.

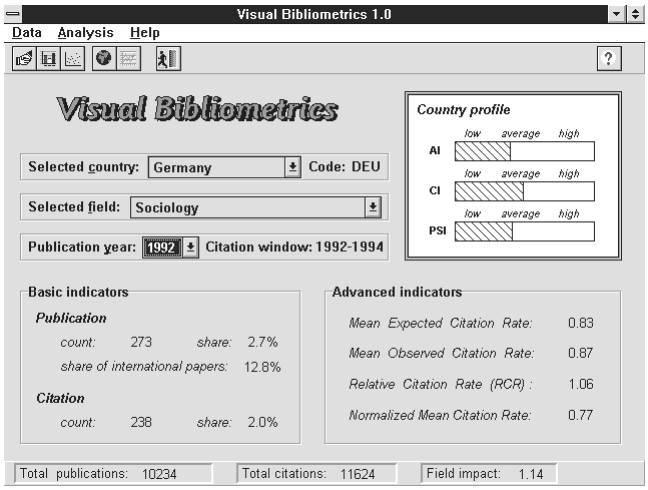



*Figure 8.4 The main window presenting bibliometric "standard" indicators and a concise country profile*

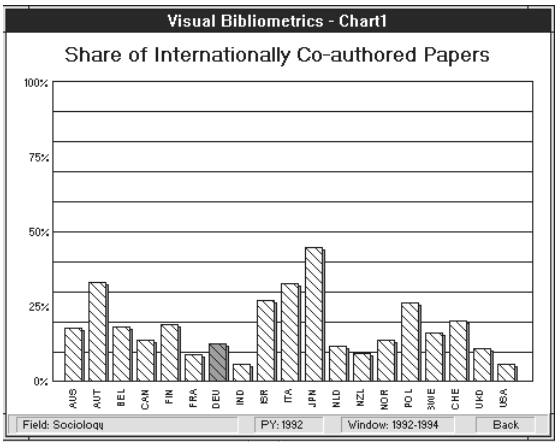



*Figure 8.5 Bar diagrams show the extent of international scientific collaboration in the selected field*

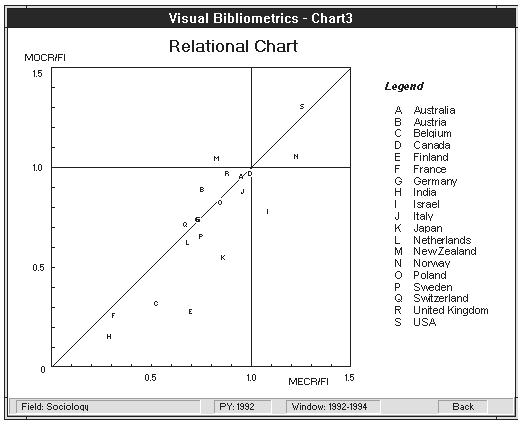



*Figure 8.6 The relational chart shows the plot of observed vs. expected citation rate in the selected field*

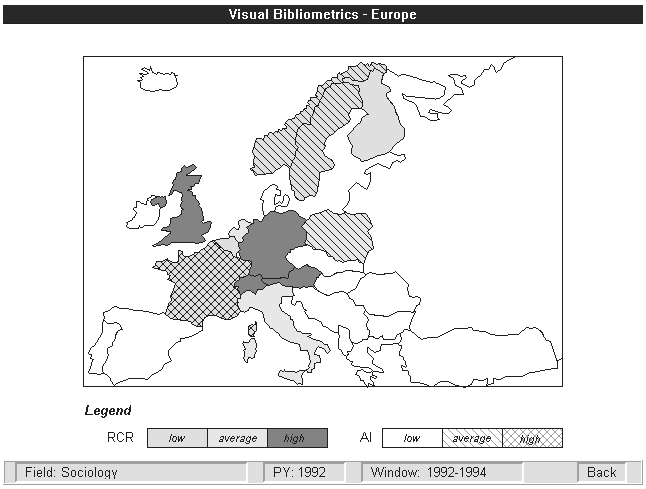



*Figure 8.7 The "bibliometric atlas" visualises the most important publication activity and citation impact indicators*

## *8.2 Bibliometric Software*

In this last section, we will give four examples for bibliometric software; two of them are freeware, the other two packages are commercial products.

### *The Bibliometrics Toolbox*

The Bibliometrics Toolbox was the first bibliometric programme package. It has been created by *Terrence A. Brookes* to assist bibliometricians in preparing statistics from their downloaded data. It comprises a set of computer programs written in Turbo Pascal that measure the bibliometric aspects of a literature such as Bradfordity, productivity ranks, degree of clustering, and indices of concentration. *Bibliometrics Toolbox* has been reviewed by *McLain* (1990), and is freeware available from the FTP address.

### *Data view bibliometric software for analysis of downloaded data*

Data view is commercial software. It has been developed by the *Centre de Recherche Rétrospective de Marseille* (CRRM,) at the Faculté Saint Jérome in Marseille (France). The software does not provide a new bibliometric method; it provides a bridge between information sources and the various data analysis methods. Dataview is designed as a software tool for experts of scientific and technological information processing. They can use this tool to build their own analysing techniques according to the most suitable statistic methods. In order to reach this purpose, dataview accepts various formats of information funds such as on-line databases or CD-Editions, allows the use of several types of bibliometric items in the same study and provides numerical data for various statistical techniques. Dataview provides the main necessary issues and the main necessary edition formats used for bibliometric analysis. The main features of this software are visualised in Figure 8.8. For more details, consult *Rostaing* et al. (1993).

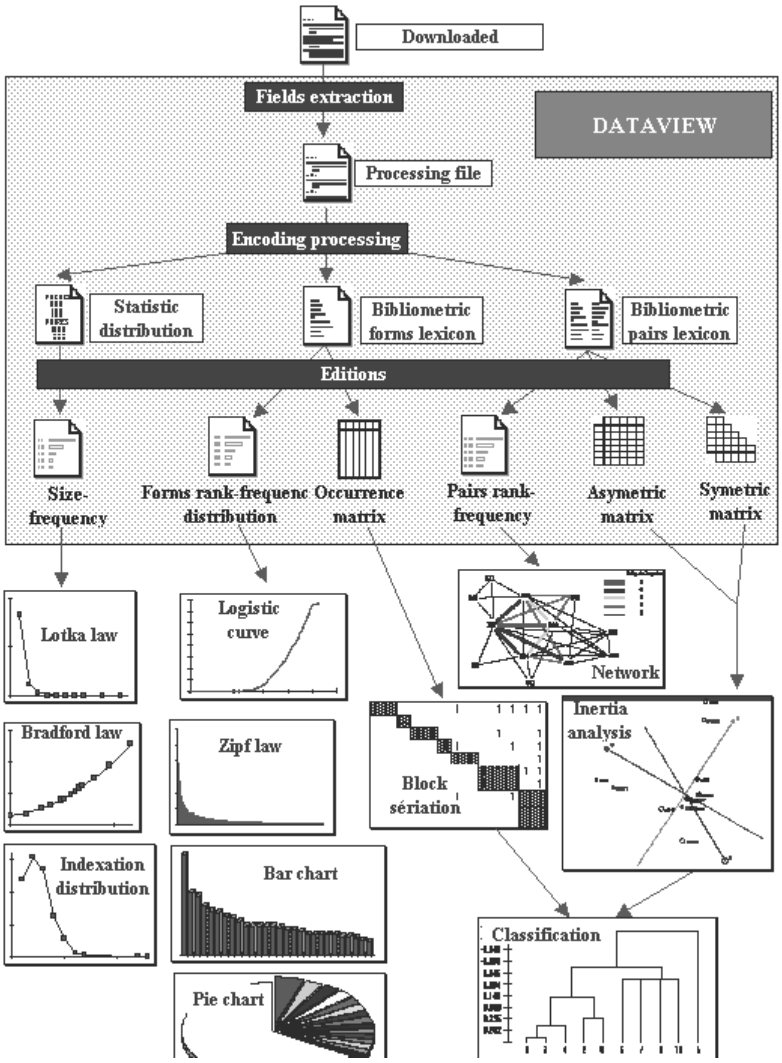



*Figure 8.8 Main features of the bibliometric software 'Dataview'*

*Bibexcel*

Bibexcel is a toolbox developed by Olle Persson, Inforsk, Umeå Univ (Sweden). This software is designed to assist a user in analysing bibliographic data, or any data of a textual nature formatted in a similar manner. The idea is to generate data files that can be imported to MS Excel, or any program that takes tabbed data records, for further processing. This toolbox includes a number of tools, some of them visible in the window and others hide behind the menus. Many of the tools can be used in combination to achieve the desired result. Figure 8.9 shows the main window of Bibexcel.

Bibexcel allows generating several maps using Multi-Dimensional-Scaling techniques. A map is made by first calculating the number of times pairs of units, for example authors, co-occur in the document records. Then the resulting co-occurrence matrix is taken as input to a Multi-Dimensional-Scaling program that finds the best fitting two-dimensional representation of the input values. The distance between units on the map is inversely proportional to the number of co-occurrences, which means that the more two units co-occur the closer they will be located on the map. The maps presented in Figures 8.10 and 8.11 are examples by the author of Bibexcel.

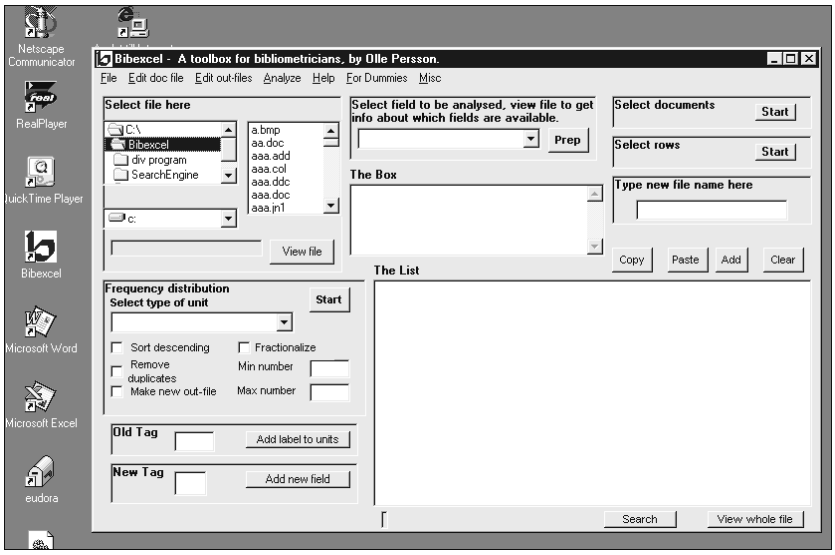



*Figure 8.9 The main window of Bibexcel*

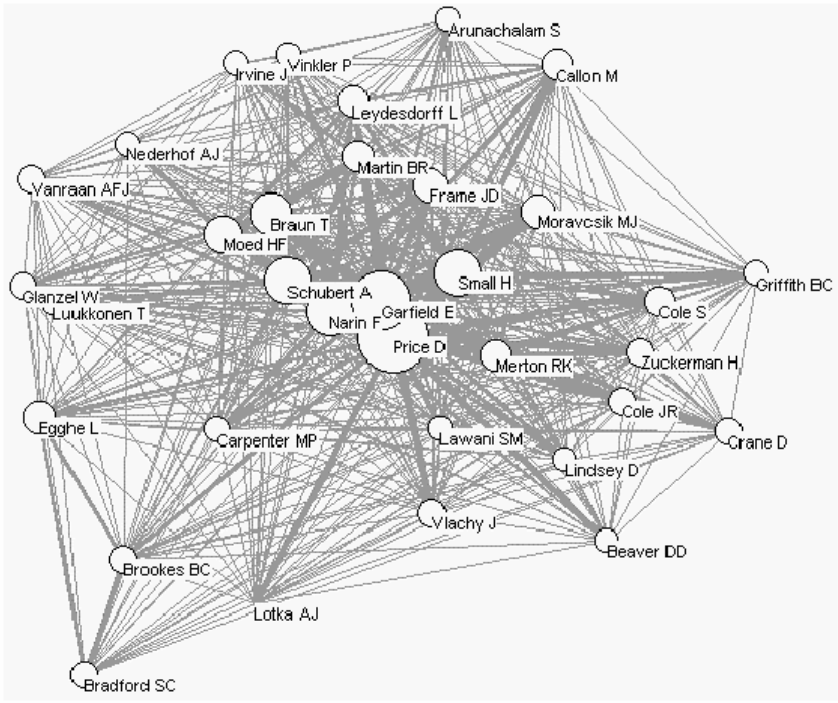



*Figure 8.10 Intellectual base of scientometrics 1978–1999. First author co-citations made by 1062 papers in the journal Scientometrics (redrawn from Persson, 2000)*

Figure 8.11 visualises first author co-citations made by 1062 papers in the journal Scientometrics. This is based on the ACA method suggested by *White* and *McCaine* (1998). The units are represented by labels, which can be coloured to show a certain attribute, for example the national origin of an author. A circle can be drawn at each node indicating its size, for example the number of papers written by an author. Lines between nodes as well as their thickness indicate the number of co-occurences. This example has been taken from *Persson* (2000).

The second example given in Figure 8.11 shows the collaboration among main institutions in Finland on the basis of co-publications. The location of institutions on the map is estimated by applying a Multi-Dimensional-Scaling algorithm to a collaboration matrix. This example has been taken from *Persson* et al. (2000).
Bibexcel is freeware, and can be downloaded from the author's homepage.

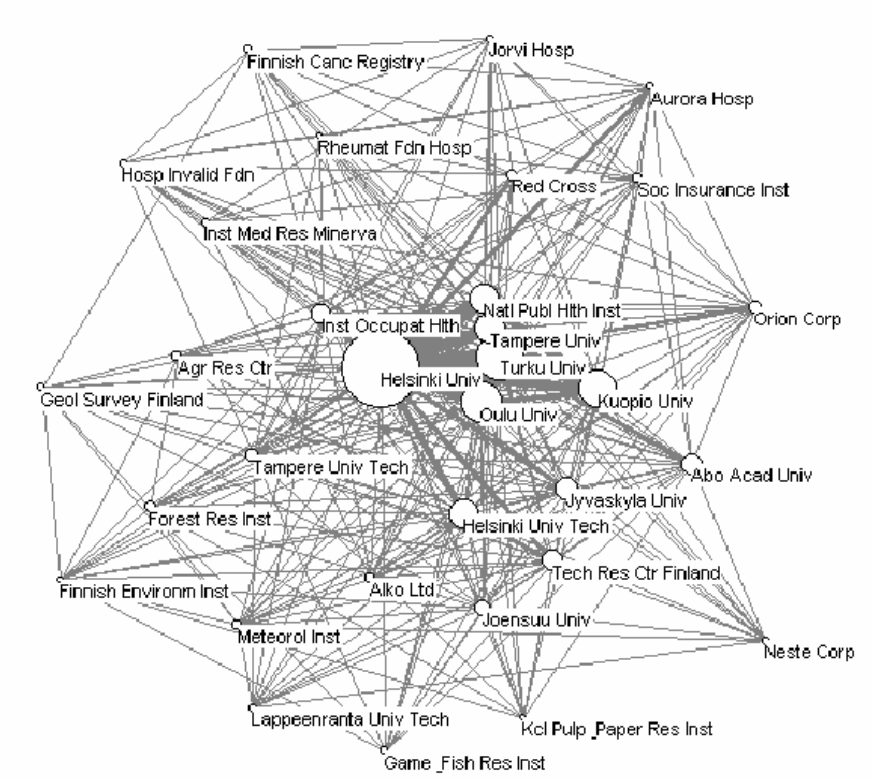



*Figure 8.11 Collaboration among main institutions (according to Persson et al., 2000)*

*BibTechMon*$^{TM}$

BibTechmon is the last example for bilbliometric software. It is commercial product developed at the Austrian Research Centres Seibersdorf (Austria). BibTechMon can be used for the documentation and structuring of external information from patent and bibliographic databases, from the internet or other external or internal sources. The software has the following main features. It supports the process of building a literature database as well as designing knowledge maps and analysis of the content by

- reading in documents for building up an internal database
- automatic indexing for identification and classification of relevant key-terms
- calculation of knowledge maps
- analysis of the maps using an interactive surface or browser
- direct access to the documents in a database
- database search functions

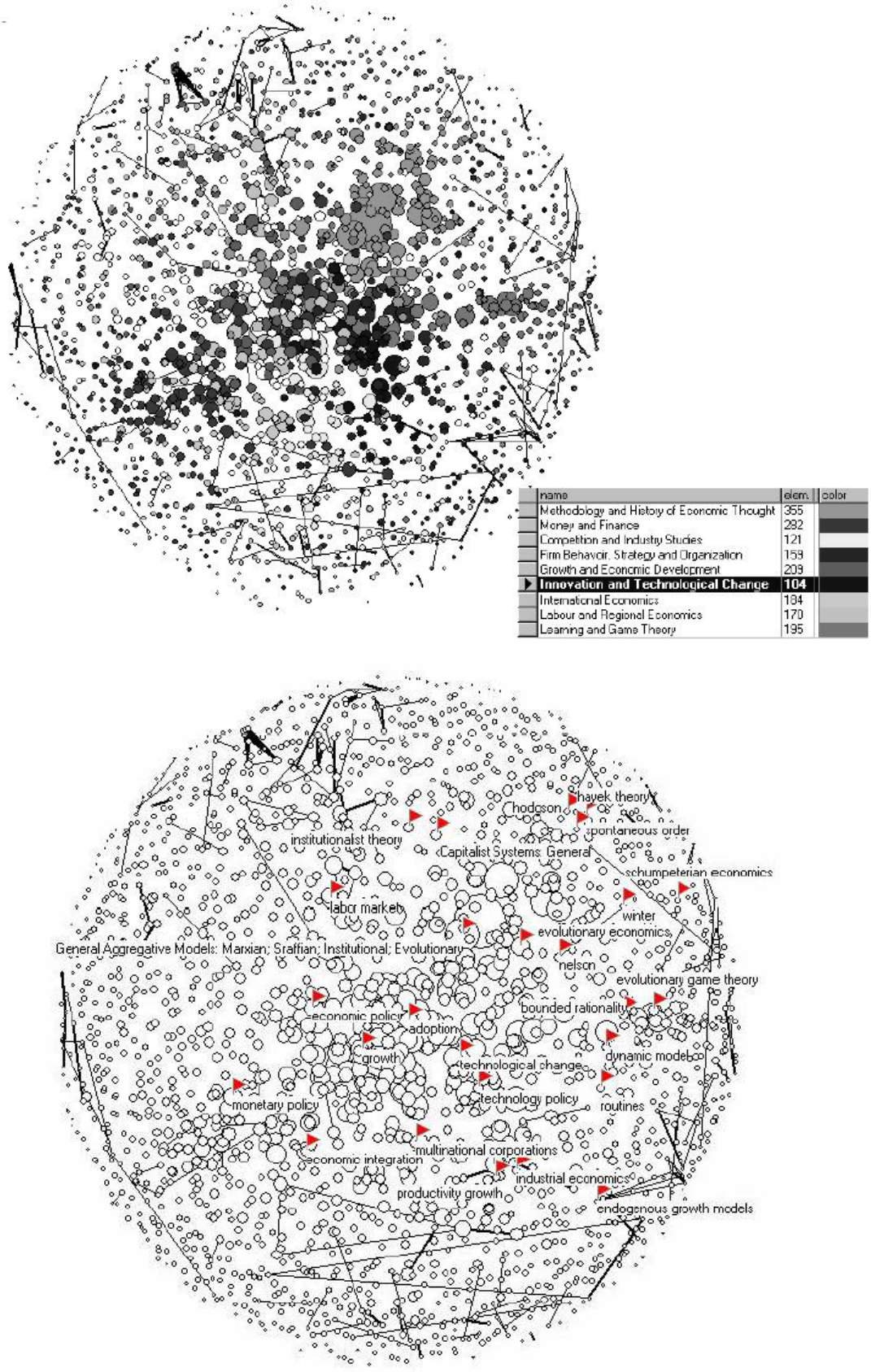



*Figure 8.12 Knowledge map of key terms (top) and location of various key terms (bottom) forevolutionary economics*

Figure 8.12 taken from *Dachs* et al., 2001 presents knowledge map and location of key terms for 'evolutionary economic'. The authors have retrieved literature from the EconLit database. Data have been cleaned up by removing distortions based on the context sensitive longest-match principle and a phrase recognition algorithm (*Widhalm* et al., 1999). This automatic indexing module has been applied to titles, sources and abstracts of each record. The mapping algorithm for the co-word analysis is based on the *Jaccard Index* as similarity measure. Maps are then generated applying an MDS algorithm. The underlying mathematical-statistical procedure is decribed by *Kopsa* and *Seibel* (1998). Further information about this software and its fields of application can be found, for instance, on *Clemens Widhalm's* home page.

## REFERENCES


AKSNES, D. W., A macro-study of self-citations. *Scientometrics*, **56** (2), 2003, 235-246.

ALLISON, P., Inequality of Scientific Productivity, *Social Studies of Science,* **10**, 1980, 163-179.

ALLISON, P., D. PRICE, B. GRIFFITH, M. MORAVCSIK, J. STEWART, Lotka's Law: A Problem in its Interpretation and Application. *Social Studies of Science*, **6,** 1976, 269-276.

ASAI, I., Adjusted Age Distribution and Its Application to Impact Factor and Immediacy Index, *Journal of the American Society for Information Science*, **32**, 1981, 172-174.

BEAVER, D. DEB., ROSEN, R., Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship, *Scientometrics,* **1**, 1978, 65-84.

BEAVER, D. DEB., ROSEN, R., Studies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French elite, *Scientometrics,* **1**, 1979, 133-149.

BEAVER, D. DEB., Reflections on scientific collaboration (and its study): past, present, and future *Scientometrics*, **52** (3), 2001, 365-377.

BOOKSTEIN, A., Informetric Distributions.
Part I: Unified Overview, *JASIS*. **41**(5), 1990, 366-375.
Part II: Resilience to Ambiguity, *JASIS*, **41**(5), 1990, 376-386.

BONZI, S., SNYDER, H. W., Motivations for citation - A comparison of self citation and citation to others, *Scientometrics*, **21** (2), 1991, 245-254.

BJÖRNEBORN, L., INGWERSEN, P., Perspectives of webometrics. *Scientometrics*, **50** (1), 2001, 65-82.

BRAAM, R R, MOED, H. F, VAN RAAN, A. F. J., Mapping of science by combined co-citation and word analysis, part II: Structural aspects. *JASIS,* **42**, 1991, 233-51.

BRAUN, T., GLÄNZEL, W., SCHUBERT, A., *Scientometric Indicators. A 32 Country Comparison of Publication Productivity and Citation Impact*. World Scientific Publishing Co. Pte. Ltd., Singapore * Philadelphia, 1985.

BRAUN, T., BÚJDOSÓ, E., SCHUBERT, A., *Literature of analytical chemistry: A scientometric evaluation,* CRC Press, Inc., Boca Raton, Florida, 1987.

BRAUN, T., GLÄNZEL, W., On a Source of Error in Computing Journal Impact Factors, *Chemical Intelligencer*, January, 1995, 31-32.

BRAUN, T. GLÄNZEL, W., SCHUBERT, A., Publication and cooperation patterns of the authors of neuroscience journals, *Scientometrics,* **51** (3), 2001, 499-510.

BRADFORD, S.C., Sources of Information on Specific Subjects, *Engineering*, **137**, 1934, 85-86.

BROAD W. J., *Study Finds Publicly Financed Science Is a Pillar of Industry*, New York Times Science Times Section, 13 May, 1997.

CALLON, M., COURTIAL, J-P., TURNER, W., BRAIN, S., From translations to problematic networks: An introduction to co-word analysis, *Social Science Information*, **22**, 1983, 191-235.

CALLON, M., LAW, J., RIP, A. *Mapping of the dynamics of science and technology*. London: McMillian, 1986.

CHAPMAN, A. J., Assessing Research: Citation-Count Shortcomings. *The Psychologist: Bulletin of the British Psychological Society*, **8** (8), 1989, 339-341.

COURTIAL, J-P., A co-word analysis of scientometrics. *Scientometrics* **31** (3), 1994, 251-260.

COZZENS, S. E., What do citations count? The rhetoric-first model. *Scientometrics*, **15,** 1989, 437-447.

CRONIN, B., The need for a theory of citation, *Journal of Documentation* **37**, 1981, 16-24.

DACHS, B., ROEDIGER-SCHLUGA, T., WIDHALM, C., ZARTL, A., *Mapping Evolutionary Economics. A Bibliometric Analysis*, Paper prepared for the EMAEE 2001 Conference, University of Economics and Business Administration, Vienna, 2001.

DEBACKERE, K. (red.). *Vlaams Indicatorenboek Wetenschap, Technologie, Innovatie*, AWI en IWT publicatie, 2003, in press.

DE BRUIN, R. E., MOED, H. F., Delimitation of Scientific Subfields Using Cognitive Words from Corporate Addresses in Scientific Publications, *Scientometrics*, **26** (1), 1993, 65-80.

DE LEEUW, J., VAN DER HEIJDEN, P., *Quasi-correspondence Aanalysis*. Univ Leiden. RR-85-19, 1985.

BROOKES, B. C. The growth, utility, and obsolescence of scientific periodical literature. *Journal of Documentation*, **26**, 1970, 283-94.

BUJDOSÓ, E., *Bibliometria és tudománymetria*, Országos Széchényi Könyvtár Könyvtártudományi és Módszertani Központ - MTA Könyvtára, Budapest, 1986.

BURRELL, Q.L., Stochastic Modelling of the First Citation Distribution. *Scietometrics*, **52**, 2001, 3-12.

COURTIAL, J.P., *Introduction à la scientometrie*, Anthropos, Paris, 1990.

EGGHE L., ROUSSEAU, R., A characterization of distributions which satisfy Price's law and consequences for the laws of Zipf and Mandelbrot. *Journal of Information Science*, 12, 1986, 193-197.

EGGHE, L., ROUSSEAU, R., *Introduction to Informetrics. Quantitative Methods in Library*, Documentation and Information Science, Elsevier, Amsterdam, 1990.

EGGHE L., On the influence of growth on obsolescence. *Scientometrics*, **27** (2) 1993, 195-214

EGGHE, L., Price Index and its relation to the mean and median reference age. *JASIS*, **48** (6), 1997, 564-573.

EGGHE, L., A heuristic study of the first-citation distribution. *Scientometrics*, **48** (3), 2000 345-359.

FRAME, J.D. 1977, Mainstream research in Latin America and the Caribbean, *Interciencia*, 2, 143-148.

GARFIELD, E. "The 'Obliteration Phenomenon' in Science – and the Advantage of Being Obliterated!" *Essays of an Information Scientist,* **2**. Philadelphia: ISI Press, 1977, 398.

GARFIELD, E., *Citation Indexing: Its Theory and Applications in Science, Technology and Humanities*. New York: Wiley, 1979.

GARFIELD, E., Uses and Misuses of Citation Frequency. *Essays of an Information Scientist*, **8**. Philadelphia: ISI Press, 1986a, 407.

GARFIELD, E., The Evolution of Physical Chemistry to Chemical Physics, *Current Contents*, 3, 1986b, 3-12.

GARFIELD, E., Essays of an Information Scientist. Vol. 1-13. ISI Press, Philadelphia, 1977-1991.

GARFIELD, E., What is a Citation Classic? *Current Contents* (appears weekly)

GELLER, N.L., Citation Influence Methodology of Pinski and Narin, *Information Processing and Management*, **14**, 1978, 93-95.

GLÄNZEL, W., Stochasztikus akció-reakció folyamotok és alkalmazásuk a tudománymetriában. (Stochastic action-reaction processes and their application to scientometric problems). Doctoral thesis. University Budapest, 1983.

GLÄNZEL, W., SCHUBERT, A., Price Distribution. An Exact Formulation of Price's "Square Root Law". *Scientometrics,* **7** (3-6) (1985) 211-219.

GLÄNZEL, W. On some stopping times on citation processes. From theory to indicators. *Information Processing and Management,* **28** (1), 1992a, 53-60.

GLÄNZEL, W., *Publication Dynamics and Citation Impact: A Multi-Dimensional Approach to Scientometrics Research Evaluation*. In P. Weingart, R. Sehringer, & M. Winterhager (Eds.), Representations of Science and Technology. Proceedings of the International Conference on Science and Technology Indicators, Bielefeld, Federal Republic of Germany, 10-12 June 1990 Leiden: DSWO Press, 1992b, pp. 177-188.

GLÄNZEL, W., SCHUBERT, A., Some Facts and Figures on Highly Cited Papers in the Sciences, 1981-1985, *Scientometrics*, **25** (3), 1992, 373-380.

GLÄNZEL, W., SCHOEPFLIN, U., A Stochastic Model for the Ageing Analyses of Scientific Literature, *Scientometrics*, **30** (1), 1994, 49-64.

GLÄNZEL, W., SCHUBERT, A., Predictive Aspects of a Stochastic Model for Citation Processes, *Information Processing & Management*, **31** (1), 1995, 69-80.

GLÄNZEL, W., SCHOEPFLIN, U., A Bibliometric Study on Ageing and Reception Processes of Scientific Literature. *Journal of Information Science*, **21** (1), 1995, 37-53.

GLÄNZEL, W., CZERWON, H.J., A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level, *Scientometrics* **37**, 1996, 195-221.

GLÄNZEL, W., The need for standards in bibliometric research and technology, S*cientometrics*, **35** (2), 1996, 167-176.

GLÄNZEL, W. , SCHOEPFLIN, U., A Bibliometric Study of Reference Literature in the Sciences and Social Sciences, *Information Processing and Management,* **35**, 1999, 31-44.

GLÄNZEL, W., SCHUBERT, A., CZERWON, H. J., An Item-by-item Subject Classification of Papers Published in Multidisciplinary and General Journals Using Reference Analysis, *Scientometrics*, **44**, 1999, 427-439.

GLÄNZEL, W., Science in Scandinavia: A Bibliometric Approach, *Scientometrics*, **48** (2), 2000, 121-150.

GLÄNZEL, W., National Characteristics in International Scientific Co-authorship, *Scientometrics*, **51** (1), 2001, 69-115.

GLÄNZEL, W., MOED, H. F., (Eds.), Journal impact measures: their role in research policy and scientific information management., *Selected papers of the Special Day Session at the 8th International Conference on Scientometricsand Informetrics,* held in Sydney (Australia) on 17 July, 2001, dedicated issue of the journal *Scientometrics*, **53** (2), 2002, 169-279.

GLÄNZEL, W., MEYER, M., *Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations in the Triple Helix*, Paper presented at the 4th Triple Helix Conference, held in Copenhagen (Denmark) – Lund (Sweden), on 6-9 November, 2002.

GLÄNZEL, W., THIJS, B., SCHLEMMER, B., *A bibliometric approach to the role of author self-citations in scientific communication*, lecture to be presented at the 9th International Conference on Scientometrics and Informetrics, August 25-29, 2003, Beijing, China

GLÄNZEL, W., *On some principle differences between citations and sitation links*. NERDI lecture delivered at NIWI, KNAW, Amsterdam, on Februrary 13, 2003. Available at www.niwi.knaw.nl/nerdi/lectures/glanzel.pdf (Updated version of a paper presented at the Sixth Nordic Workshop on Bibliometrics, Stockholm, October 4-5, 2001.)

GORKOVA, V.I., Informetrics, *Informatics*, **10**, VINITI, Moscow, 1988.

GROSS, P.L.K., GROSS, E.M., College Libraries and Chemical Education*, Science*, **66**, 1927, 385-389.

GOFFMAN, W., An epidemic model in an open population, *Nature*, 205, 1965, 831.

GOFFMAN W., NEVILL, V.A., Generalization of epidemic theory. *Nature*, 204, 1964, 225.

GUPTA, R.C., On the Mean Residual Life Function in Survival Studies. In: C. Teil et al. (eds.), *Statistical Distributions in Scientific Work*, Vol. 5, D. Reidel, Dordrecht-Holland, 1981, 327-334.

ХАЙТУН, С. Д., Наукометрия. Состоение и перспективы, Изд. Наука. Москва, 1983.

ХАЙТУН, С. Д., Проблемы количественого анализа науки, Изд. Наука. Москва, 1989.

HICKS, D., Limitations of co-citation analysis as a tool for science policy, *Social Studies of Science,* **17**, 1987, 295-316.

HICKS, D., 360 Degree Linkage Analysis, *Research Evaluation*, **9**, 2000,133-143.

HOFER GEE, H., NARIN, F., *An Analysis of Research Publication Supported by NIH*, 1973-1980. NIH Program Evaluation Report. U.S. Department of Health and Human Services. NIH Publication No. 86-2777, 1986.

JANSZ, C.N.M., LEPAIR, C., Bibliometric invisibility of technological advances, In*:* P. Weingart, R. Sehringer, M. Winterhager (Eds.), *Presentations of Science and Technology*, DSWO Press, Leiden, 1992, 315-326.

JANSZ, C.N.M., The citation gap for applicable sciences and the search for new technology indicators, *Proceedings of the 7th International Conference on Scientometrics and Informetrics*, held on 5-8 July 1999, in Colima (Mexico), Universidad de Colima, 1999, 234-243.

KARLIN, S., TAYLOR, H. M., *A first course in stochastic processes*. Academic Press, Inc., 1975.

KATZ, J.S., Bibliometric standards: Personal experience and lessons learned. *Scientometrics*, **35** (2), 1996, 193--197.

KÄRKI, R., KORTELAINEN, T. *Introduktion till bibliometri*. Översättning och bearbetning av kristina Eriksson & Sara von Ungern-Sternberg. Nordinfo publikation 41, Helsingfors, 1998.

KESSLER, M. M., Bibliographic coupling between scientific papers, *American Documentation,* **14**, 1963, 10-25.

KRETSCHMER, H., (1992a), Significance of the logarithm of number of publications for the measurement of social stratification in coauthorship networks. - In: *Informetrics 91/92*. Select papers from the 3rd International Conference on Informetrics. 9-12 August 1991. Ravichandra Rao, I. K. (Ed.), Bangalore: Sarada Ran-ganat-han Endowment for Library Science, 1992, 289-331.

KOPCSA**,** A., SCHIEBEL**,** E., Science and Technology Mapping: A New Iteration Model for Representing Multidimensional Relationships, *JASIS*, **49** (1), 1998, 7-17.

KRETSCHMER, H., (1992b), The adaption of coauthorship networks to changing conditions of the research process. - In: *Science and Technology in a Police Context*. Select Proceedings of the Joint EC-Leiden Conference on Science and Technology Indicators. Leiden 23-25 October 1991, van Raan, A.F.J., de Bruin, R.E., Moed, H.F., Nederhof, A.F., Tijssen, R.W.J., Leiden (Eds). DSWO Press University Leiden 1992. 280-288.

KRETSCHMER, H., Coauthorship networks of invisible colleges and institutional communities. *Scientometrics*, **30** (1) 1994, 363-369.

KRETSCHMER, H., ROUSSEAU, R., Author inflation leads to a breakdown of Lotka's law. *JASIST*, **52**, 8, 2001, 610-614.

LAUDEL, G., What do we measure by co-authorships? *Research Evaluation*, **11** (1), 2002, 3-15.

LAW, J., WHITTAKER, J., Mapping acidification research: A test of the co-word method, *Scientometrics*, **23** (3), 1992, 417-461.

LEIMKUHLER, F. F. The Bradford distribution, *Journal of Docurnentation*, **23**, 1967, 197-207.

LEWISON, G., DAWSON, G., ANSERSON, J., The behaviour of biomedical scientific authors in acknowledgind their funding sources. *Proceedings of the 5th International Conference on Scientometrics and Informetrics*, held in River Forest, Illinois, June 7-10, Learned Information Inc., Medford, 1995, 255-263.

LEWISON, G., DAWSON, G., The effect of funding on the outputs of biomedical research. *Scientometrics,* **41**(1-2), 1998, 17-27.

LEYDESDORFF, L., Theories of Citation?, *Scientometrics* **43** (1), 1998, 5-25

LINDSEY, D., Corrected Quality Ratio: A Composite Index of Scientific Contribution to Knowledge, *Social Studies of Science*, **8**, 1978, 349-354.

LINE, M.B., The structure of social science literature as shown by a large-scale citation analysis. *Social Science Information Studies*, **1**, 1981, 67-87.

LOTKA, A.J., The Frequency Distribution of Scientific Productivity, *J. Washington Acad. Sci.*, **16**, 1926, 317-323.

LUUKKONEN, T., PERSSON, O., SILVERTSEN, G., Understanding patterns of international scientific collaboration, *Science, Technology & Human Values*, **17,** 1992, 101-126.

LUUKKONEN, T. TIJSSEN, R. J. W. PERSSON, O. SILVERTSEN, G., The measurement of international scientific collaboration, *Scientometrics*, **28,** 1993, 15-36.

MANDELBROT , B. New methods in statistical economics, *Journal of Political Economy*, **71**, 1963, 421-440.

MCLAIN, J.P., Bibliometrics toolbox, *JASIS* 41 (1), 1990, 70-71.

MACROBERTS, M. H., MACROBERTS. B. R., Problems of citation analysis: A critical review. *JASIS,* **40** (5), 342-349

MARDIA, K.V., KENT, J.T., BIBBY, J.M., *Multivariate Analysis*. Academic Press Inc. (London) LTD, 1979.

MARSHAKOVA, I. V., System of connections between documents based on references (as the Science Citation Index), *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2*, No. 6, 1973, 3-8 (in Russian).

MOED, H. F., Bibliometric Measurement of Research Performance and Price's Theory of Differences among the Sciences, *Scientometrics*, **15** (5-6), 1989, 473-483.

MOED, H.F., VAN LEEUWEN, TH. N., Improving the accuracy of the Institute for Scientific Information's Journal Impact Factor. *JASIS*, **46,** 1995, 461-467.

MOED, H.F., VAN LEEUWEN, TH. N., Impact Factors Can Mislead. *Nature*, **381**, 1996, 186.

MOED, H.F., VAN LEEUWEN, TH. N., REEDIJK, J., A Critical analysis of the Journals Impact Factors of *Angewandte Chemie* and *the Journal of the American Chemical Society*: Inaccuracies in Published Impact Factors Based on Overall Citations Only, *Scientometrics,* **37**, 1996, 105-115.

MOED, H.F., VAN LEEUWEN, TH. N., REEDIJK, J., A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation*, **54**, 1998, 387-419.

MOED, H.F., VAN LEEUWEN, TH. N., REEDIJK, J. A. Towards appropriate indicators of journal impact. *Scientometrics*, **46** (3), 1999, 575-589

НАЛИМОВ, В. В., МУЛЬЧЕНКО, З.М., Наукометрия, Изд. Наука. Москва, 1969.

NARIN, F., *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*, Computer Horizons, Inc., Washington, D.C., 1976.

NARIN, F., HAMILTON, K. S., OLIVASTRO, D., The Increasing Linkage between U.S. Technology and Public Science, *Research Policy*, **26**, (3), 1997, 317-330.

NEDERHOF, A.J., MEIJER, R.F., MOED, H.F., VAN RAAN, A.F.J., Research Performance Indicators for University Departments - A Study of an Agricultural University, *Scientometrics*, **27** (2), 1993, 157-178.

NOYONS, E.C.M., VAN RAAN, A.F.J., GRUPP, H., SCHMOCH, U., Exploring the science and technology interface: Inventor-author relations in laser medicine research, *Research Policy*, **23**, 1994, 443-457.

PERITZ, B. C., Are Methodological Papers More Cited than Theoretical or Empirical Ones? The Case of Sociology, *Scientometrics*, **5** (4) 1983, 211-218.

PERSSON, O., A tribute to Eugene Garfield – Discovering the intellectual base of his discipline, *Current Science*, **79** (5), 10 September 2000, 590-591.

PERSSON, O., LUUKKONEN, T., HÄLIKKÄ, S., *A Bibliometric Study of Finnish Science*, Working Papers No. 48/00, VTT, Group for Technology Studies, Espoo, 2000.

PERSSON, O., GLÄNZEL, W., DANELL, R., *Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies*, paper to be presented at the 9th International Conference on Scientometrics & Informetrics, August 25-29, 2003, Beijing, China

PINSKI, G., NARIN, F., Citation Influence for Journal Aggregates of Scientific Publications, *Information Processing and Management*, **12**, 1976, 297-312.

PRICE, D. DE SOLLA, *Little Science, Big Science*, Columbia Univ. Press, New York, 1963.

PRICE, D. SE SOLLA Citation measures of hard science, soft science, technology, and non-science. In: Nelson, C.E., Pollak, D.K. (Eds.): *Communication among Scientists and Engineers* (Heat, Lexington, Mass., 1970, 1-12.

PRICE, D. SE SOLLA., A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science,* **27***,* 1976, 292–306.

PRICE, D. SE SOLLA., The analysis of square matrices of scientometric transaction, *Scientometrics*, **3**, (1), 1981, 55-63.

PRITCHARD, A., Statistical bibliography or bibliometrics? *Journal of Documentation* 24, 1969, 348-349.

RAVICHANDRA RAO, I. K., Quantitative Methods for Library and Information Science. Wiley-Eastern. New Delhi, 1983.

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

ROSTAING, H., NIVOL, W., QUONIAM, L., LA TELA, A., Le logiciel bibliometrique Dataview et son application comme outil d'aide a l'evaluation de la concurrence, *Revue Française de Bibliometrie*, N°12, 1993, 360-387.

ROUSSEAU, R., Double exponential model for first-citation processes. *Scientometrics*, **30** (1), 1994, 213-227.

ROUSSEAU, R., Sitations: an exploratory study, *Cybermetrics*, **1** (1), 1997. Paper 1. http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

SCHUBERT, A., ZSINDELY, S., BRAUN, T., Scientometric analysis of attendance at international scientific meetings. *Scientometrics*, **5** (3), 1983, 177-187.

SCHUBERT, A., GLÄNZEL, W., Mean Response Time. A new indicator of journal citation speed with application to physics journals. *Czech. J. Phys.,* **B 36** (1986) 121-125.

SCHUBERT, A., TELCS, A., Estimation of the publication potential in 50 U.S. states and in the District of Columbia based on the frequency distribution of scientific productivity. JASIS, **40** (4), 1989, 291-297.

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

SCHUBERT, A., BRAUN, T., World Flash on Basic Research: International Collaboration in the Sciences, 1981-1985, *Scientometrics*, **19**, 1990, 3-10.

SCHUBERT, A., BRAUN, T., Reference-Standards for Citation Based Assessments, *Scientometrics*, **26** (1), 1993, 21-35.

SEN, S. K., GAN, S. K., A mathematical extension of the idea of bibliographic coupling and its applications, *Annals of Library Science and Documentation*, **30** (1983), 78-82.

SHARABCHIEV, J.T., Comparative analysis of two methods of cluster analysis of bibliographic references, *Nauchno-Tekhnicheskaya Informatsiya, Seriya 2*, No. 4, 1988, 25-28 (in Russian).

SICHEL, H. S., A bibliometric distribution which really works. *JASIS,* **36***,* 1985, 314-321.

SICHEL, H. S., Anatomy of the Generalized Inverse Gaussian-Poisson Distribution with special applications to bibliometric studies. *Information Processing and Management* **28** (1), 1992, 5-18.

SMALL, H., Co-citation in the scientific literature: A new measure of the relationship between two documents, *Journal of the American Society for Information Science*, **24,** 1973, 265-269.

SMALL, H., Cited documents as context symbols, *Social Studies of Science*, **8**, 1978, 327-240.

SMALL, H., Citation context analysis, *Progress in Communication Science*, **3**, 1982, 287-310.

SMALL, H., A general framework for general large-scale maps of science in two or three dimensions: The SciViz system. *Scientometrics*, **41**(1-2), 1998, 125-133.

SMART, J.C., ELTON, C.F., Consumption Factor Scores of Psychology Journals, *Scientometrics*, **4**, 1982, 349-360.

SMITH, L. C., Citation Analysis. *Library Trends*, **30** (1), 1981, 85.

TIJSSEN, R. J. W., DE LEEUW, J., VAN RAAN, A. F. J., Quasi-correspondence analysis on Scientometric Transaction matrices, *Scientometrics*, 11, 1987, 351-366.

TIJSSEN, R.J.W., DE LEEUW, J., VAN RAAN, A.F.J., A Method for Mapping Bibliometric Relations Based on Field-Classifications and Citations of Articles. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88*, Elsevier Science Publishers B.V., Amsterdam, 1988, p. 279-292.

TODOROV, R, Condensed Matter Physics Journal, *Scientometrics*, **5**, 1983, 291-301.

TODOROV, R, Representing Canadian geophysics: A bibliometric approach. In: L. Egghe, R. Rousseau (Eds.), *Informetrics 89/90*, Elsevier Science Publishers B.V., 1990, 291-307. (Proceedings of the 2nd International Conference on Bibliometrics, Scientometrics and Informetrics, held in London (Canada), 4-7 July 1989)

TODOROV, R., GLÄNZEL, W., Computer Bibliometrics for Journal Classification. *Information Processing and Management*, **26** (5) (1990) 673-680.

TOMER, C., A Statistical Assessment of Two Measures of Citation: the Impact Factor and the Immediacy Index, *Information Processing and Management*, **22**, 1986, 251-258.

TURNER, W., CHARTRON, G., LAVILLE, F., MICHELET, B. Packaging information for peer review: new co-word analysis techniques. In: Van Raan, A., (Ed), *Handbook of quantitative studies of science and technology*. 291-323, Amsterdam: North Holland, 1988.

VANELS, W.P., JANSZ, C.N.M., LEPAIR, C., The citation gap between printed and instrumental output of technological research - the case of the electron-microscope, *Scientometrics* **17** (5-6), 1989, 415-425.

VAN LEEUWEN, TH. N., VISSER, M. S., SPRUYT, E., MOED, H. F., Bibliometrische Studie van de Faculteiten Wetenschappen, Geneeskunde en Farmaceutische Wetenschappen aan de Universiteit Antwerpen (1981-1998), CWTS Rapport 2001-01, 2001, Leiden.

VAN LEEUWEN, TH. N., H.F. MOED, J. REEDIJK. *JACS* Still Topping *Angewandte Chemie*: Beware of Erroneous Impact Factors. *Chemical Intelligencer*, July*,* 1997, 32-36.

VAN RAAN, A.F.J. (Ed.), *Handbook of Quantitative Studies of Science and Technology,* North-Holland, Amsterdam, 1988.

VERBEEK, A., DEBACKERE, K., LUWEL, M., ANDRIES, P., ZIMMERMANN, E., DELEUS, F., Linking science to technology: Using bibliographic references in patents to build linkage schemes Scientometrics, **54** (3), 2002, 399 - 420

VINKLER, P., A kémia tudományos kommunikációs rendszereinek tudománymetriai vizsgálata (Scientometric analysis of scientific communication systems in chemistry). Dr. Acad. Thesis, Budapest, 2002.

VLACHY, J., Physics papers most cited in 1961-1982 and their successive citation, *Czechoslovak Journal of Physics*, **B 36** (7), 1986, 887-890.

VLADUTZ, G., COOK, J. Bibliographic coupling and subject relatedness, In: 1984: *Challenges to an Information Society,* Proceedings of the 47th ASIS Annual Meeting (compiled by B. Flood, J. Witiak, T. H. Hogan), White Plains: Knowledge Industry Publications, 1984, 204-207.

WALLACE, D.P., The relationship between journal productivity and obsolescence. *Journal of the merican Society for Information Science*, **37,** 1986, 136-145.

WATSON, G.S., WELLS, W.T., On the Possibility of Improving the Mean Useful Life of Items by Eliminating those with Short Lives. *Technometrics*, **3**, 1961, 281-298.

WEINSTOCK, N. Citation indexes, In Kent A. (Ed.). *Encyclopedia of Library and Information Science*, New York: Marcel Dekker, **5**, 1971, 16-41.

WESTNEY, L. C. H., Historical rankings of science and technology: A citationist perspective, *The Journal of the Association for History and Computing*, **1** (1), June 1998

WIDHALM, C*., et al.*, Konzeptive Entwicklung eine Einlesesystems und einer Strategie zur automatischen Schlagwortgenerierung, OEFZS-S-0051, 1999, confidential.

YANOVSKY, V.I., Citation Analysis Significance of Scientific Journals, *Scientometrics*, **3**, 1981, 223-233.

ZIPF, G. K., *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.

# APPENDIX

## RECOMMENDED READINGS

### I. Fundamental works


1. Price, D. de Solla 1961, *Science since Babylon*, Yale Univ. Press, New Haven.
2. Price, D. de Solla 1963, *Little Science, Big Science*, Columbia Univ. Press, New York.
3. Pritchard, A. 1969, Statistical bibliography or bibliometrics? *Journal of Documentation* 24, 348-349.
4. van Raan, A.F.J. (Ed.) 1988, *Handbook of Quantitative Studies of Science and Technology,* North-Holland, Amsterdam.
5. Ziman, J. 1984, *An Introduction to Science Studies - The Philosophical and Social Aspects of Science and Technology.* Cambridge University Press, Cambridge.


### II. Bibliometric methods developed for science studies


6. Callon, M., Courtial, J.P., Turner, W.A. and Bauin, S. 1983, From translations to problematic networks: An introduction to co-word analysis, *Social Science Information* 22, 191-235.
7. Carpenter, M. P. and Narin, F. 1981, The adequacy of the Science Citation Index (SCI) as an indicator of international scientific activity, *Journal of the American Society for Information Science* 32, 430-439.
8. Garfield, E. and Welljams-Dorof, A. 1992, Citation data: their use as quantitative indicators for science and technology evaluation and policy-making, *Science and Public Policy* 19, 321-327.
9. Glänzel, W. and Czerwon, H.J. 1996, A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level, *Scientometrics* 37, 195-221.
10. Glänzel, W., Schubert, A., Czerwon, H. J., An Item-by-item Subject Classification of Papers Published in Multidisciplinary and General Journals Using Reference Analysis, *Scientometrics*, 44, 1999, 427-439.
11. Hicks, D. 1987, Limitations of co-citation analysis as a tool for science policy, *Social Studies of Science,* 17, 295-316.
12. Irvine, J. and Martin, B. R. 1989, International comparisons of scientific performance revisited, *Scientometrics* 15, 369-392.
13. Moed, H.F., de Bruin, R. E. and van Leeuwen, Th. N. 1995, New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications, *Scientometrics*, 33, 381-422.
14. Schubert, A. and Braun, T. 1986, Relative indicators and relational charts for comparative assessment of publication output and citation impact, *Scientometrics*, 9, 281-291.
15. Schubert, A. and Braun, T. 1992, Standards for citation based assessments, *Scientometrics*, 26, 21-35.

16. Small, H.G. 1973, Co-citation in scientific literature. A new measure for the relationship between publications. *JASIS,* 24, 265-269.

17. Small, H.G. and Griffith, B. C. 1974, The structure of scientific literatures I: Identifying and graphing specialities, *Science Studies* 4, 17-40.

**III. Selected comparative studies based on science indicators**

18. Glänzel, W. 1996, Scientometric Indicators Datafiles. A bibliometric approach to social sciences. National research performances in 6 selected social science areas, 1990-1992, *Scientometrics,* 35, 291-307

19. Martin, B. R., Irvine, J. and Isard, P. 1990, Trends in UK government spending on academic and related research: a comparison with FR Germany, France, Japan, the Netherlands and USA, *Science and Public Policy* 17, 3-13.

20. Miquel, J. F., Ojasoo, T., Okubo, Y., Paul, A., Doré, J. C. 1995, World science in 18 disciplinary areas: Comparative evaluation of the publication patterns of 48 countries over the period 1981-1992, *Scientometrics,* 33, 149-167

21. Glänzel, W., Science in Scandinavia: A Bibliometric Approach, *Scientometrics*, 48, 2000 121-150 (Correction: *Scientometrics*, 49, 2000, 357)

22. Schubert, A., Glänzel, W., Braun, T. 1989 World flash on basic research: Scientometric datafiles. A Comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields, 1981-1985, *Scientometrics,* 16, 3-478.

**IV. Research in industries**

23. Frumau, C.C.F. 1992, Choices in R&D and business portfolio in the electronics industry: What the bibliometric data show, *Research Policy*, 21, 97-124.

24. Godin, B. 1996, Research and the practice of publication in industries, *Research Policy*, 25, 587-606.

25. Grupp, H. 1990, On the supplementary functions of science and technology indicators, *Scientometrics*, 19, 447-472.

26. Hicks, D., Ishizuka, T., Keen, P. and Sweet, S. 1994, Japanese corporations, scientific research and globalization, *Research Policy*, 23, 375-384.

27. Narin, F. and Rozek R.P. 1988, Bibliometric analysis of U.S. pharmaceutical industry research performance, *Research Policy*, 17, 139-154.

**V. International scientific collaboration**

28. deB. Beaver, D., Rosen, R. 1978, Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship, *Scientometrics,* 1, 65-84

29. deB. Beaver, D., Rosen, R., 1979, Studies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French elite, *Scientometrics,* 1, 133-149

30. Braun, T., Gómez, Isabel, Méndez, Aida, Schubert, A., 1992, World flash on basic research: International co-authorship patterns in physics and its subfields, 1981-1985, *Scientometrics,* 24, 181-200.

31. Gómez, I., Fernández, M. T. and Méndez, A. 1995, Collaboration patterns of Spanish scientific publications in different research areas and disciplines, In: *Proceedings of the Biennial Conference of the International Society for Scientometrics and Informetrics* (ed. by M.E.D. Koenig and A. Bookstein), Learned Inf., Medford, NJ, pp. 187-196.

32. Hicks, D. and Katz, J.S. 1996, Science policy for a highly collaborative science system, *Science and Public Policy*, 23, 39-44.

33. Lewison, G. and Cunningham, P. 1991, Bibliometric studies for the evaluation of trans-national research, *Scientometrics*, 21, 223-244.

34. Luukkonen, T., Persson, O., Silvertsen, G. 1992, Understanding patterns of international scientific collaboration, *Science, Technology & Human Values*, 17, 101-126

35. Luukkonen, T., Tijssen, R. J. W., Persson, O., Silvertsen, G. 1993, The measurement of international scientific collaboration, *Scientometrics*, 28, 15-36.

36. Moed, H.F., de Bruin, R.E., Nederhof, A.J. and Tijssen, R.J.W. 1991, International scientific co-operation and awareness within the European Community: problems and perspectives, *Scientometrics*, 21, 291-311.

37. Narin, F., Stevens, K. and Whitlow, E.S. 1991, Scientific co-operation in Europe and the citation of multinationally authored papers, *Scientometrics*, 21, 313-323.

38. Vinkler, P. 1993, Research Contribution, Authorship and Team Cooperativeness. *Scientometrics*, 26, 213-230

**VI. Science in developing countries**

39. Arunachalam, S. and Garg, K. C. 1986, Science on the periphery - a scientometric analysis of science in the ASEAN countries, *Journal of Information Science*, 12, 105-117.

40. Arvanitis, R., Gaillard, J. (Eds) 1992, *Science Indicators for Developing Countries*, ORSTOM, Paris

41. Gibbs, W.W. 1995, Lost science in the Third World, *Scientific American*, 273, 76-83.

42. Moravcsik, M.J. 1985, Applied scientometrics: An assessment methodology for developing countries, *Scientometrics,* 7, 165-176

43. Sancho, R. 1992, Misjudgements and shortcomings in the measurement of scientific activities in less developed countries, *Scientometrics*, 23, 221-233.

44. Stolte-Heiskanen, V. 1986, Evaluation of scientific performance on the periphery, *Science and Public Policy*, 13, 83-88.

**Additional recommended literature**

Braun, T., Bújdosó, E., Schubert, A., 1987, *Literature of analytical chemistry: a scientometric evaluation,* CRC Press, Inc., Boca Raton, Florida.

Tijssen, R. J. W., 1992, *Cartography of science: scientometric mapping with multi-dimensional scaling techniques*. DSWO Press, Leiden University

Noyons, E.C.M., 1999, *Bibliometric mapping as a science and research management tool.* DSWO Press, Leiden University

Wouters, P., 1999, *The citation culture*, PhD Thesis, Private edition.

White, H.D., McCain, K.W., 1989. Bibliometrics. *Annual Reviews of Information Science and Technology*, 24, 119-186.

Garfield, E., 1998, From citation indexes to infometrics: Is the tail now wagging the dog? *Libri*, 48, 67-80.

Borgman, C.L., 1990, *Scholarly communication and bibliometrics*, Sage.

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.

Persson, O., 1994, The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science*, 45, 31-38.