Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1174

Published online 20 December 2010 in Wiley Online Library

Influence of Database Mistakes on Journal Citation Analysis: Remarks on the Paper by Franceschini and Maisano, QREI (2010)

Fiorenzo Franceschini*† and Domenico Maisano

This short note contains some remarks on a recent bibliometric survey about some of the major scientific journals in the field of Quality Engineering/Management (*Qual. Reliab. Engng. Int.* 2010; 26(6):593–604). In particular, thanks to Professor Woodall's precious indication, it has been freshly noticed that some results in the original work are biased by mistakes in the bibliometric databases (in this case Google Scholar). After a careful examination and correction of biased data, a synthetic analysis of the typical mistakes of bibliometric databases is presented, focussing the attention on the importance of using robust bibliometric indicators. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: bibliometric indicators; citation analysis; Hirsch-index; robust indicator; database mistake; false reference

1. Introduction

This short note concerns a paper recently published in the *Quality and Reliability International* journal by the authors[1]. Briefly, this paper presented a bibliometric analysis of 12 major journals in the field of Quality Engineering/Management, from the point of view of some bibliometric indicators (see Table I). It is worth noticing that, when large-scale evaluations of the scientific production are performed (e.g. over hundreds or even thousands of publications), bibliometric indicators seem to be the only practicable instrument[2, 3].

Indicators used in the original analysis are, respectively,

- *Hirsch (h) index for a journal*, defined as the number such that, for the group of articles published by the journal in a precise time period (e.g. one year), *h* articles received at least *h* citations while the others received no more than *h* citations[4, 5]. Figure 1 shows the original *h* profiles for the journals of interest, in 20 consecutive years.
- *Total number of citations* (C), defined as the number of citations received up to the moment of the analysis by the journal issue(s) published in a specific period (e.g. in one year). Figure 2 shows the original C profiles for the journals of interest, in 20 consecutive years.
- *h-spectrum*, defined as the distribution representing the *h* values associated with the authors (and coauthors) of a specific journal, considering a specific publication period. This indicator provides an image of the journal author population in a precise time period[6].

For a detailed explanation of these indicators, we refer the reader to the original article[1].

Citation statistics were collected using the Google Scholar (GS) database. It was decided to use this database (i) because of the greater coverage with respect to other databases (such as Web of Science (WoS) or Scopus) and (ii) since it is free and can be easily accessed through Publish or Perish© (PoP) or other *ad hoc* software applications, specially designed for citation analysis with GS[7]. Indicators were calculated taking into account the citations accumulated up to the moment of the original analysis (June 2009).

That being said, the decision of writing this short note was taken after receiving from Professors Woodall and Montgomery some comments on the data represented in Figure 2 [8, 9]. In detail, the comments concern the C profile related to 'Technometrics' (TM), which looks rather nervous, with many peaks that often fall beyond the upper limit of the vertical axis scale.

Politecnico di Torino, Dipartimento di Sistemi di Produzione ed Economia dell'Azienda (DISPEA), Corso Duca degli Abruzzi 24, 10129 Torino, Italy
*Correspondence to: Fiorenzo Franceschini, Politecnico di Torino, Dipartimento di Sistemi di Produzione ed Economia dell'Azienda (DISPEA), Corso Duca degli Abruzzi 24, 10129 Torino, Italy.
†E-mail: fiorenzo.franceschini@polito.it

Table I. List of the 12 Quality journals selected for the analysis

Journal name	Acronym	Publisher	Indexed by Thomson Scientific
IIE Transactions (on Quality and Reliability Engineering)	IIETR	Taylor & Francis	Yes
International Journal of Quality and Reliability Management	IJQRM	Emerald	No
Journal of Quality in Maintenance Engineering	JQME	Emerald	No
Journal of Quality Technology	JQT	ASQ	Yes
Managing Service Quality	MSQ	Emerald	No
Quality Engineering	QE	ASQ	No
Quality Management Journal	QMJ	ASQ	No
Quality Progress	QP	ASQ	No
Quality and Quantity	QQ	Springer	Yes
Quality and Reliability Engineering International	QREI	Wiley	Yes
Technometrics	TM	ASQ	Yes
Total Quality Management & Business Excellence	TQM	Taylor & Francis	No

Journals are sorted in alphabetical order with respect to the journal acronym. Adapted from Reference[1].

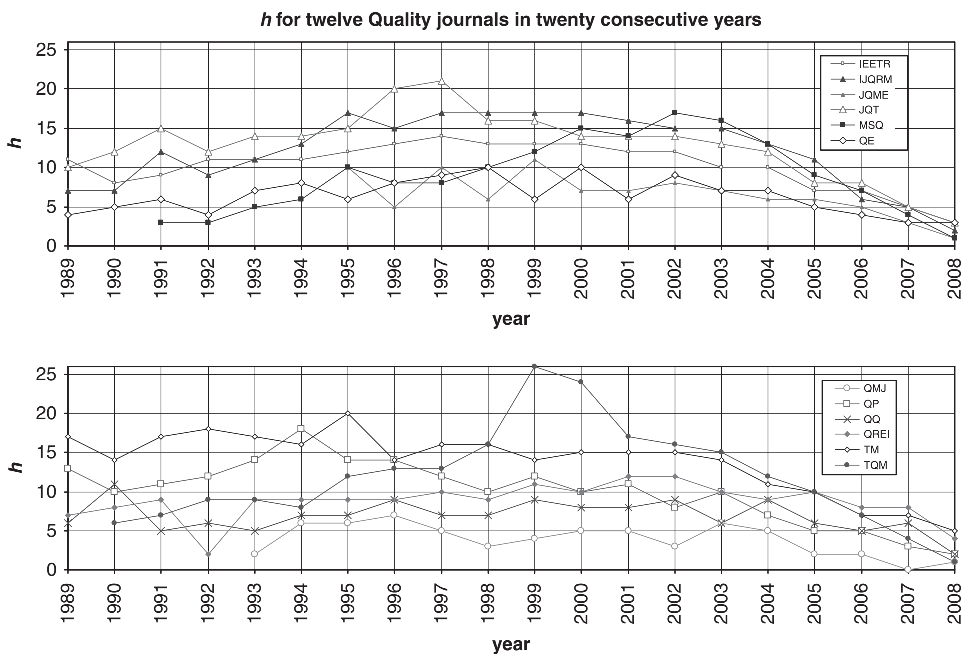

Figure 1. *h* values for the 12 Quality journals (see Table I), in 20 consecutive years (from 1989 to 2008). Values are calculated taking into account the citations accumulated up to the moment of the original analysis (June 2009). For the purpose of readability, journal profiles are first sorted in alphabetical order with respect to the journal acronyms and then divided into two groups of six each. Adapted from Reference[1]

In the initial analysis, it was noticed that in 1992, 2000 and 2004, C values were inflated by a small number of 'big hit' articles with a huge number (several hundreds) of received citations. These 'big hits' are the focal point of this short note. We give Professor Woodall merit for noticing that some of them are false references, originated from database mistakes. Specifically

1. The surprisingly high number of citations assigned to some papers published in 2000 is due to the fact that, in this year, TM republished some of the best papers from his history (e.g. important papers of Mallows, Roberts, Nelson, etc.) and these were all counted as 2000 papers by GS (see the PoP screenshot in Figure 3(a))[10]. A proof of this mistake is that almost all the citations received by those papers are prior to 2000 (Figure 4 shows, for example, some of the citations associated with the very highly cited Hoerl and Kennard's paper[11], republished in 2000). Probably, testing for the condition that the citing year must be equal to or larger than the publication year of the cited document could filter out a large number of false matches like these.
2. Most of the citations in 2004 are associated with a book[12], which was reviewed in a 2004 TM issue (see the PoP screenshot in Figure 3(b)). Book citations were counted in error as citations of a TM paper.

In a recent analysis it has been discovered a database error of the same nature as the second one, for 2003. Specifically, most of the citations received by TM in 2003 are associated with another book[13], reviewed in a 2003 TM issue (see the PoP screenshot

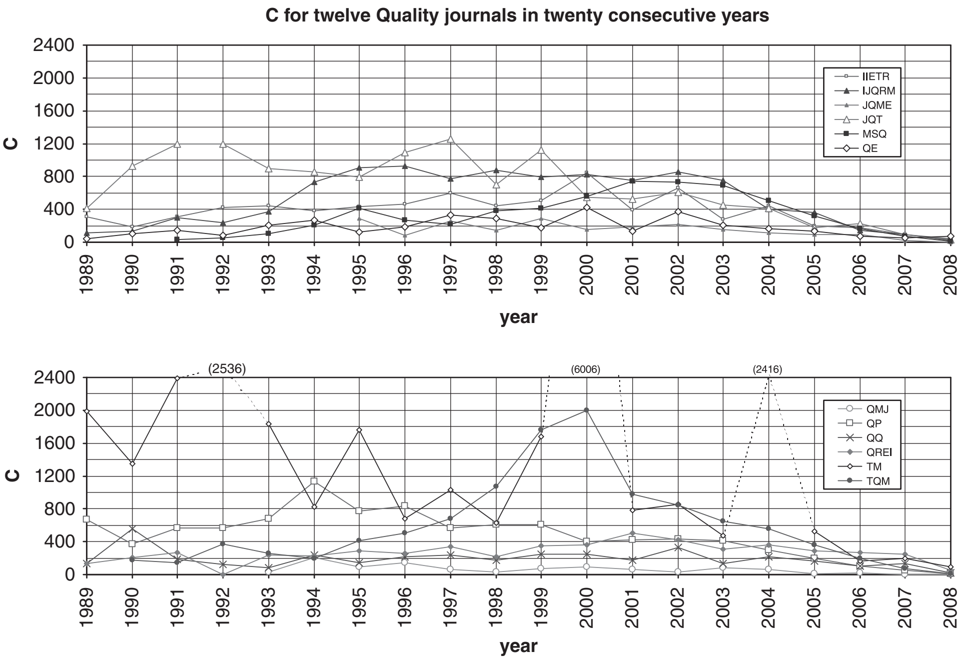

Figure 2. C values for the 12 Quality journals (see Table I), in 20 consecutive years (from 1989 to 2008). Values are calculated taking into account the citations accumulated up to the moment of the original analysis (June 2009). For the purpose of readability, journal profiles are first sorted in alphabetical order with respect to the journal acronyms and then divided into two groups of six each. Adapted from Reference[1]

in Figure 3(c)). This case is even more striking than the previous one, since 10 005 out of the 10 626 total citations received are related to this 'false reference'. Moreover, 2003 citations reported in the original survey (in Figure 2) are not affected by the latter anomaly, probably because the 'false reference' occurred later.

The verification of the previous database mistakes has been carried out in a recent analysis (August 2010), compared to the original one (June 2009). The results will be discussed in more detail in the subsequent sections.

The remainder of this short note is organized into two sections. Section 2 illustrates the new C and *h* profiles related to the 12 journals of interest, repeating the analysis and removing database mistakes discussed before. Section 3 makes some general comments on the common mistakes of bibliometric databases and the possible remedies to limit their negative effect.

2. New *h* and *C* profiles

Figures 5 and 6 report, respectively, the new C and *h* profiles related to the scientific journals in Table I, after removing the false references discussed in Section 1. It can be noticed that these new profiles are not very dissimilar from those of the original analysis, although slightly higher. This depends on the fact that, in the about 14-month period between the original analysis (June 2009) and the recent update (August 2010), the corresponding publications have accumulated new citations. Analysing the new TM profile, it can be noticed that 2000 and 2004 peaks are no longer present. On the other hand, there are two more outlying years (i.e. 1991 and 1993). Regarding the remaining 11 journals, other anomalies due to database errors have not been identified. The only exception is a new peak of 'Total Quality Management & Business Excellence' (TQM) in 2000. In this year, TQM has a relatively high number of articles with many citations and a significant portion of them have been accumulated in the time elapsed between the original analysis and the recent update.

3. Remarks on database mistakes

3.1. *Sensitivity of the indicators in use*

Using the words of Professor Montgomery, database mistakes noticed by Professor Woodall are evidence of how tricky the citation business can be. In particular, mistakes like these may significantly influence some non-robust bibliometric indicators. For example, let us consider the average number of citations per paper (CPP) related to TM. Being not very robust, CPP is subject to large fluctuations between biased values (in presence of database mistakes) and corrected values (after removing database mistakes). For the purpose of example, Table II reports the values of CPP and other common indicators—i.e. the total number of publications (P), C and *h*—for TM, in the three years influenced by the database mistakes (2000, 2003 and 2004).

Examining Table II, it can be noticed that the biased C and CPP values are 1–2 orders of magnitude higher than the corrected ones. On the other hand, differences between biased and corrected *h* values are much smaller. This is an additional confirmation

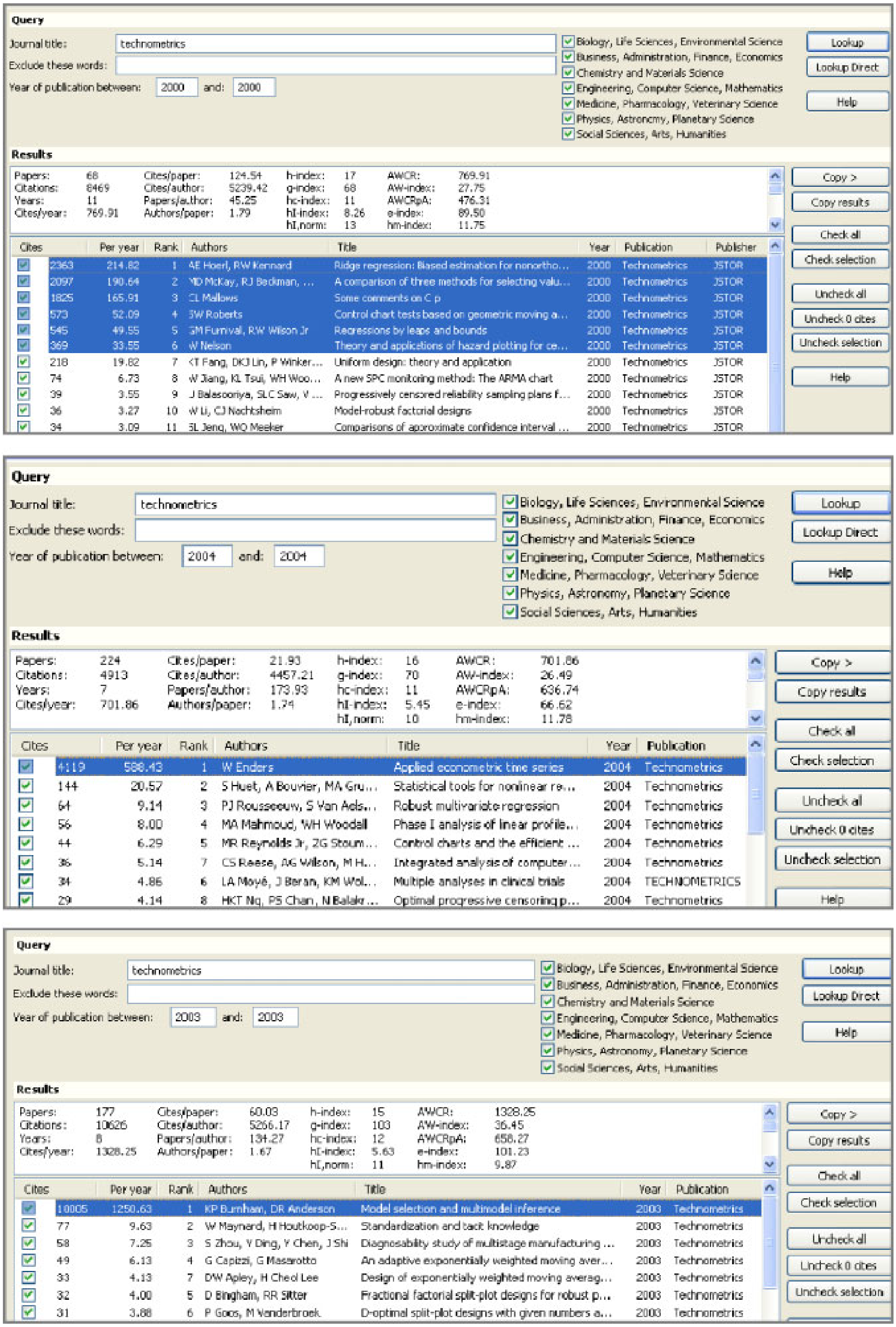

Figure 3. (a) Portion of the PoP screenshot listing the papers published in TM in 2000. Most of the 'big hit' papers (including those highlighted) are some of the best papers from TM history, republished for the 40th anniversary of the journal. (b) Portion of the PoP screenshot listing the papers published in TM in 2004. Most of the citations (4119 out of 4913) are incorrectly associated to an Enders' book[12] (highlighted), which was reviewed in a 2004 TM issue. (c) Portion of the PoP screenshot listing the papers published in TM in 2003. Most of the citations (10 005 out of 10 626) are incorrectly associated with a Burnham and Anderson's book[13] (highlighted), which was reviewed in a 2003 TM issue. Values are calculated taking into account the citations accumulated up to the moment of the recent update (August 2010)

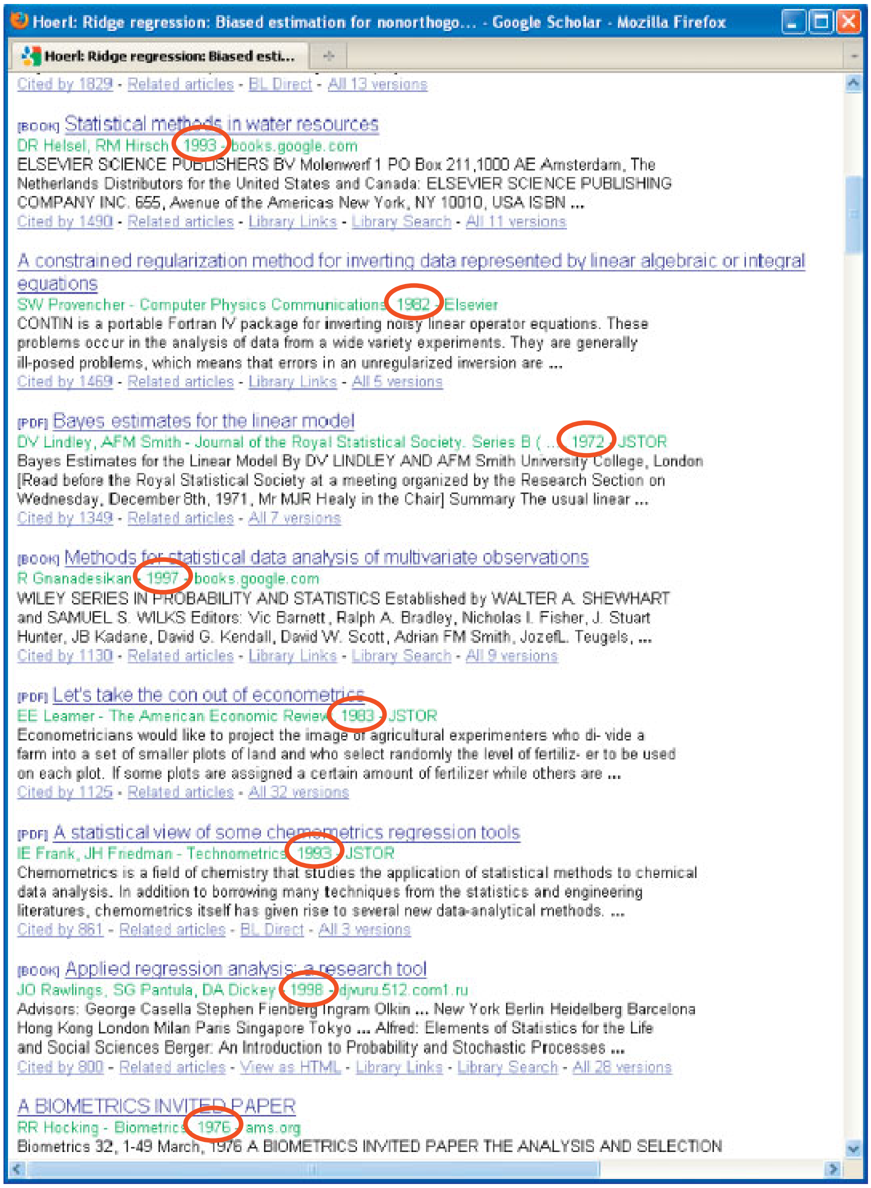

Figure 4. GS screenshot related to the papers citing the Hoerl and Kennard's article[11], republished in TM in 2000. It can be seen that most of the citations come from papers prior to 2000 (the corresponding publication dates are circled)

that *h*-index is a robust indicator, since it is not much influenced by the number of citations received by the most cited papers[14, 15]. For this reason, in many situations *h* is preferred to other indicators such as those based on the average number of CPP—including the ISI journal Impact Factor, with its false impression of precision conveyed by the three decimal points[16]. This is one of the reasons why *h* has been used in our survey, not only to compare scientific journals on the basis of the received citations, as shown in Figures 2 and 5, but also to construct the journal *h*-spectrum, that is to say an indicator of the level of the scientific reputation of the journal authors.

3.2. *Common database mistakes and problems*

Referring to the previous discussion, we take the opportunity to list some of the most common errors in bibliometric databases. For more information see the relevant literature[17–20].

1. *Limited coverage*. WoS and Scopus cover thousands of peer-reviewed academic journals from every field. However, this is still only a selected set of the journals. According to many bibliometrists, coverage for many fields such as Social Sciences, Computer Science or Engineering Science is not sufficient. The fact that more than half of the journals investigated in our survey are not indexed by ISI Thompson (see Table I), and therefore not included in WoS, is emblematic. On the other hand, GS includes citations to books, book chapters, dissertations, working papers, conference paper and journal articles published in non-ISI and Open-Access journals. For this reason, GS database probably provides a more comprehensive picture of the

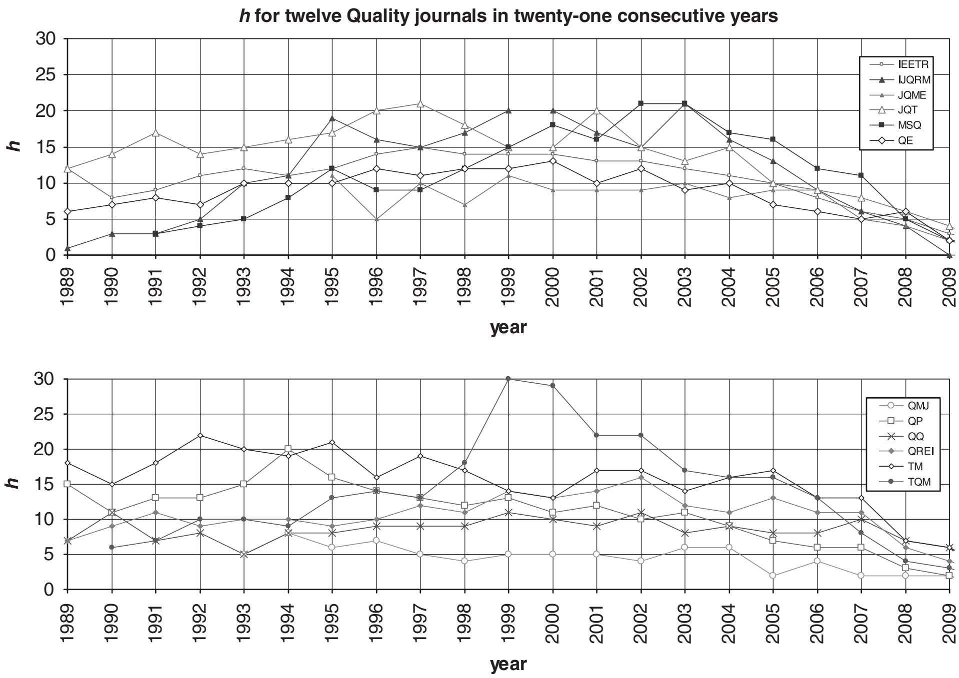

Figure 5. *h* values for the 12 Quality journals (see Table I), in 21 consecutive years (from 1989 to 2009). Values are calculated taking into account the citations accumulated up to the moment of the recent update (August 2010). For the purpose of readability, journal profiles are first sorted in alphabetical order with respect to the journal acronyms and then divided into two groups of six each

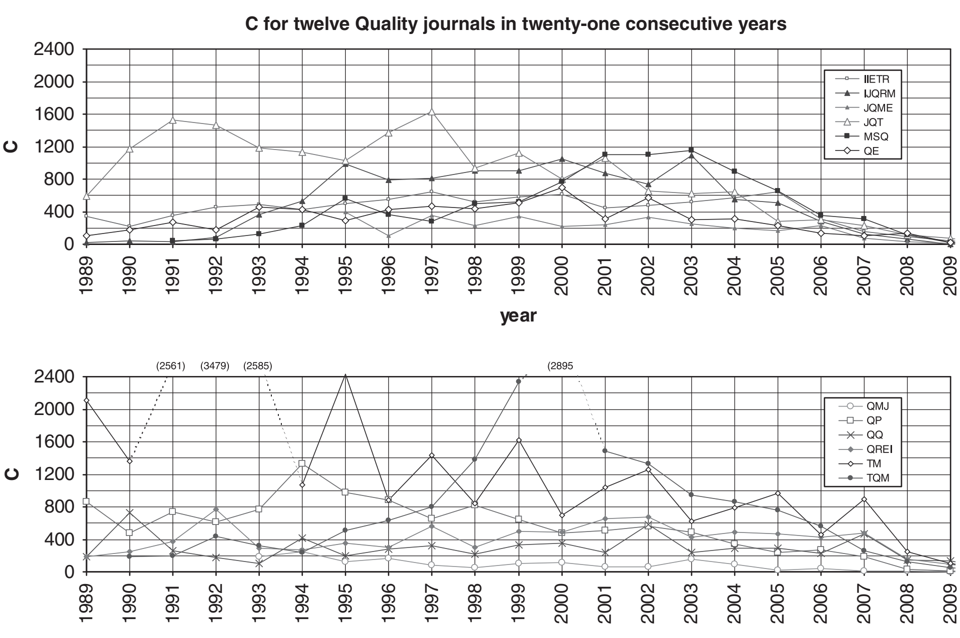

Figure 6. C values for the 12 Quality journals (see Table I), in 21 consecutive years (from 1989 to 2009). Values are calculated taking into account the citations accumulated up to the moment of the recent update (August 2010). For the purpose of readability, journal profiles are first sorted in alphabetical order with respect to the journal acronyms and then divided into two groups of six each. The profile of TM has three outlying years—precisely 1991, 1992 and 1993—while TQM profile has one outlying year (2000), falling beyond the upper limit of the vertical axis scale. The corresponding numeric values are reported in brackets. In these years, the journals have a relatively large number of articles with many citations

recent impact[7]. However, the fact that GS is still in beta testing and reluctant to declare the list of titles and documents covered tells us that this database should improve significantly before it becomes fully operational[19, 20].

2. *False references*. Results can be distorted by false references, i.e. wrong assignments of publications/citations to one author or journal, shown in the first two sections. This problem is particularly evident for GS, due to the automatic generation of the data set by scanning and parsing PDF files, to extract reference lists. This ‘strategy’ makes it possible to obtain a very high coverage and frequent update, but may sometimes sacrifice data accuracy.

Year	P	C	CPP	h	P	C	CPP	h
	Biased indicators				Corrected indicators			
	P	C	CPP	*h*	P	C	CPP	*h*
2000	68	8 469	124.5	17	62	697	11.2	13
2003	177	10 626	60.1	15	176	621	3.5	14
2004	224	4 913	21.9	16	223	794	3.6	16

Table II. Bibliometric indicators for TM in the years 2000, 2003 and 2004

Biased values are affected by some database mistakes, while corrected ones are obtained after removing these mistakes. Values are calculated taking into account the citations accumulated up to the moment of the analysis (August 2010).

3. *Duplicate records*. Bibliometric databases sometimes do not include citations to the same work that have small mistakes in their references (for example, typographical errors in titles, missing authors, authors listed in incorrect order, author names with diacritics or apostrophes, difference in the names used for the journals, and so on). The identification and summary of all duplicates is obviously crucial with regard to the quality of a citation analysis, as otherwise, relevant citations are not considered[19]. Automatic identification of duplicate records is a very challenging task that is receiving a lot of attention in computer science research[21].
4. *Author disambiguation*. When automatically compiling publication lists for authors, there exists the problem of multiple scientists having the same name[1, 6]. This problem has been tackled when selecting the authors of a specific journal, before constructing the corresponding *h*-spectrum. Several simple 'tricks'—most of them known in the literature[15]—have been resorted, in order to identify a set of authors whose names most likely correspond to only one scientist in the subject area of consideration. They are briefly presented hereafter:

(a) only scientists for which we could obtain their full *first* and last name were kept;
(b) common US and Chinese family names (such as 'Smith' or 'Chang') have been removed with the help of specific dictionaries. These cases are especially likely to be ambiguous;
(c) last names with less than five characters, such as 'Mata' or 'Tsai', have been removed. This filter removes additional names of Asian origins, for which disambiguation is a serious problem;
(d) citation statistics of authors with relative high *h*-indices, say larger than 25–30, have been manually checked due to the high risk of ambiguity;
(e) GS has been queried with full name and subject area filter, to avoid collisions with scientists in other improbable domains (i.e. Biology, Chemistry, Astronomy, etc).

Despite the more or less frequent mistakes, bibliometric databases remain essential for assessing the scientific production of scientists and/or journals and, fortunately, they seem to be more and more committed to improve and fix mistakes. However, for a sound and plausible bibliometric analysis, it is convenient to take two aspects into account[20]:

- Data preparation, data cleaning and integrating data from several sources are important to achieve useful and correct results.
- Priority should be given to robust indicators, i.e. those less affected by possible database errors. This is one of the reasons for the great diffusion of *h* and *h*-based indicators for bibliometric surveys.

References

1. Franceschini F, Maisano D. A survey of quality engineering-management journals by bibliometric indicators. *Quality and Reliability Engineering International* 2010; **26**(6):593–604. DOI: 10.1002/Qre.1083.
2. Van Raan AFJ. The Pandora's box of citation analysis: measuring scientific excellence, the last evil? *The Web of Knowledge*: *A Festschrift in Honor of Eugene Garfield*, Cronin B, Atkins HB (eds.), ASIS Monograph Series. Information Today Inc.: Medford, NJ, 2000; 301–319.
3. Orr D. Research assessment as an instrument for steering higher education—A comparative study. *Journal of Higher Education Policy and Management* 2004; **26**(3):345–362.
4. Hirsch JE. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 2005; **102**:16569–16572. DOI: 10.1073/pnas.0507655102.
5. Braun T, Glänzel W, Schubert A. A Hirsch-type index for journals. *The Scientist* 2006; **69**(1):169–173. DOI: 10.1007/s11192-006-0147-4.
6. Franceschini F, Maisano D. The Hirsch spectrum: A novel tool for analysing (academic) scientific journals. *Journal of Informetrics* 2009; **4**(1):64–73. DOI: 10.1016/j.joi.2009.08.003.
7. Harzing AW, van der Wal R. Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics* 2008; **8**(11):61–73.
8. Woodall WH. *Private Communication*, 2010.
9. Montgomery DC. *Private Communication*, 2010.
10. Kafadar K. Forty years of 'technometrics': Past, present, and future. *Technometrics* (Special 40th Anniversary Issue) 2000; **42**(1):2–4.
11. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**(1):55–67.
12. Enders W. *Applied Econometric Time Series* (2nd edn). Wiley: New York, 2004.
13. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference*: *A Practical Information-Theoretic Approach* (2nd edn). Springer: Berlin, 2002.
14. Vanclay JK. On the robustness of the h-index. *Journal of the American Society for Information Science and Technology* 2007; **58**(10): 1547–1550.
15. Henzinger M, Sunol J, Weber I. The stability of the h-index. *Scientometrics* 2010; **84**:465–479.

16. Moed HF. New developments in the use of citation analysis in research evaluation. *Archivum Immunologiae et Therapia Experimentalis* 2009; **57**:13–18.
17. Jacso P. Deflated, inflated and phantom citation counts. *Online Information Review* 2006; **30**(3):297–309.
18. Bar-Ilan J. Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics* 2008; **74**(2):257–271.
19. Bornmann L, Marx W, Schier H, Rahm E, Thor A, Daniel HD. Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by Angewandte Chemie International edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts. *Journal of Informetrics* 2009; **3**(1):27–35.
20. Bar-Ilan J. Citations to the 'introduction to informetrics' indexed by WOS, Scopus and Google Scholar. *Scientometrics* 2010; **82**(3):495–506.
21. Thor A, Rahm E. MOMA—A Mapping-based Object Matching System. *Proceedings of the Third Biennal Conference on Innovative data Systems Research*, Alisomar, CA, U.S.A., 2007; 247–248.

Authors' biographies

Fiorenzo Franceschini is a Full Professor at Politecnico di Torino, teaching Quality Engineering. Currently, he is Head of the Department of Manufacturing Systems and Business Economics (DISPEA) at the same university. He is a Senior Member of ASQ (American Society for Quality), Full Member of INFORMS (The Institute for Operations Research and Management Sciences).

Since August 1997, he is a member of the European Experts Database as an evaluator of the Research Technological Development (RTD) proposals in Industrial and Materials Technologies for the European Community.

He is author and co-author of 7 books and about 130 published papers in scientific journals, and international Conference proceedings. His current research interests are in the areas of Quality Engineering, Statistical Process Control, QFD, and Industrial Metrology. At present, he coordinates some important projects in the area of Quality Management for Public and Private organizations.

Domenico Maisano graduated cum laude in Mechanical engineering, at Politecnico di Torino, where he is currently Assistant Professor. From 2003 to 2005, he was involved in a project in the area of Quality Management for FIAT Automobiles. In 2008 he received his PhD degree in 'Systems for the Industrial Production' at Politecnico di Torino. His current research interests are industrial metrology, quality management and bibliometrics. He is co-author of 3 books and more than 30 publications on international journals and proceedings.