The citer-success-index: an indicator to select a subset of elite papers, based on citers

ISSI

2013

SCIENTOMETRICS

14th International Society of Scientometrics and Informetrics Conference

15th – 19th July 2013

Vienna, Austria

PROCEEDINGS Volume I

PROCEEDINGS OF
ISSI 2013
Vienna

VOLUME 1

14th International Society of
Scientometrics and Informetrics Conference

Vienna, Austria
15th to 20th July 2013

Editors

Juan Gorraiz, Edgar Schiebel, Christian Gumpenberger, Marianne Hörlesberger, Henk Moed

Sponsors

ASIS&T, USA
Elsevier B.V.
EBSCO Information Services, USA
Federal Ministry for Science and Research, Austria
Federal Ministry for Transport, Innovation and Technology, Austria
Information Assistant, Verein für Informationsmanagement, Vienna
ORCID, Inc.
Science-Metrix/R&D Reports
Swets Information Services
Thomson Reuters
ZSI - Centre for Social Innovation, Vienna

Printed by Facultas Verlags- und Buchhandels AG,
Stolbergasse 26, A-1050 Wien

ISBN: 978-3-200-03135-7
ISSN: 2175-1935

Organisation and Committees

Conference Chairs

Juan Gorraiz
Edgar Schiebel

Programme Chairs

Christian Gumpenberger
Marianne Hörlesberger
Henk Moed

Poster Chairs

Jacqueline Leta
Wolfgang Mayer

COMMITTEES

Local Committee

Doctoral Forum Chairs:
Ivana Roche and Christian Schlögl

Local Organising Committee:
Ulrike Felt
Peer Vries
Wolfgang Claudius Müller
Martin Fieder
Johannes Sorz
Bernard Wallner
Wolfgang Mayer
Martin Wieland
Ambros Wernisch
Manuela Kienegger
Manuela Korber
Beatrix Wepner
Maria-Elisabeth Züger
Beatrice Rath
Silvia Steinbrunner

Scientific Committee

Francis Narin
Anton Nederhof
Pentti Nieminen
Ed Noyons
Michael Ochsner
H. Peter Ohly
Carlos Olmeda-Gómez
José Luis Ortega
Maria Antonia Ovalle-Perandones
Manfred Paier
Andres Pandiella
Antonio Perianes-Rodríguez
Bluma C. Peritz
Olle Persson
Fernanda Peset
Alan Porter
Anastassios Pouris
Gangan Prathap
Junping Qiu
Luc Quoniam
Ismael Rafols
Ravichandra Rao
Emanuela Reale
Steve Reding
Ralph Reimann
John Rigby
Nicolás Robinson
Ivana Roche
Jürgen Roth
Ronald Rousseau
Jane Russell
Victor Rybachuk
Bibhuti Sahoo
Ulf Sandström
Shivappa Sangam
Elias Sanz
Kate Sapir
Andrea Scharnhorst
Edgar Schiebel
Christian Schloegl
Ulrich Schmoch
Jesper Schneider
Antoine Schoen
Torben Schubert
András Schubert
Philip Shapira
Robert Shelton
Gunnar Sivertsen
Stig Slipersæter
Henry Small
Johannes Sorz
Divya Srivastava
Marie Strahle
Cassidy Sugimoto
Yuan Sun
Mike Thelwall
Bart Thijs
Robert Tijssen
Yasar Tonta
Daniel Torres-Salinas
Metin Tunc
Peter van Den Besselaar
Nees Jan Van Eck
Thed Van Leeuwen
Bart Van Looy
Anthony Van Raan
Benjamin Vargas-Quesada
Liwen Vaughan
Peter Vinkler
Martijn Visser
Bernard Wallner
Ludo Waltman
Xianwen Wang
Beatrix Wepner
Ambros Wernisch
Howard White
Jos Winnink
Matthias Winterhager
Dietmar Wolfram
Yishan Wu
Erjia Yan
Ying Ye
Lin Zhang
Yajuan Zhao
Dangzhi Zhao
Michel Zitt
Alesia Zuccala

Preface

The 14th International Society of Scientometrics and Informetrics Conference takes place at the University of Vienna 15-19 July 2013 and is jointly organised by the University of Vienna and the Austrian Institute of Technology (AIT) under the auspices of ISSI – the International Society for Scientometrics and Informetrics.

This conference provides an international open forum for scientists, research managers, authorities and information professionals to debate the current status and advancements of informetric and scientometric theories, concepts and indicators. In addition to the traditional evaluative focus, participants will discuss practical applications in related fields such as library and information science, history of science, philosophy of science, R&D-management, etc.

This conference raises particularly the issues of new metrics (usage metrics and altmetrics) as complement to the classical citation metrics and opens the floor to discuss manifold aspects: what can really be measured with them as proxies, which could turn out to be adequate and robust indicators, and finally which reliable data sources are available to retrieve them?

The importance of this topic is underpinned by two plenary sessions. In the first one keynote speaker Johan Bollen provides an overview of social network services and analyses. In the second one old metrics are contrasted with new ones in short introductions by experts (Henk Moed, Juan Gorraiz, Victor Henning) and followed by a panel discussion with representatives from research, research management and information industry, who will shed light on the pros and cons of these indicators from their specific point of view.

The third plenary session deals with an evergreen as much as cumbersome topic, namely the methodological and ethical problems of individual-level evaluative bibliometrics. Wolfgang Glänzel and Paul Wouters will present "10 things one must not do with individual-level bibliometrics" followed by "10 things one can do with individual-level bibliometrics", both commented by Henk Moed and Gunnar Sivertsen.

The ISSI conference is certainly one of the world's largest international conferences devoted to this field, as is illustrated by the large number of 338 submissions received this year. 912 authors are affiliated to organisations located in 42 countries from all over the world. The top three contributing countries are China (149), Spain (129) and the USA (101). Chile, Cuba, Malaysia, Sri Lanka and Ukraine are represented by at least one author, too.

All contributions were evaluated by at least three reviewers of the International and Local Committees. Thereof only 145 (107 full papers and 38 research in progress papers) could be accepted for oral presentations. 36 sessions run in parallel thrice a day in groups of four covering the gamut from "citation analysis" to "open access". In addition, 107 posters are shown in two dedicated poster sessions.
All oral presentations and posters can be found in the conference proceedings.

Moreover, four tutorials either deal with several mapping tools (like e.g. "Sci2" and "Citespace") or address the unification issue of organizations, whereas four pre-conference workshops focus on information retrieval, topic extraction methods, standards for classifications, and bibliometric analysis for funding agencies. The pre-conference day is complemented by a doctoral forum.

By organising the 14th International Conference in Vienna we hope not only to extend the tradition of the ISSI conferences as one of the most important international meeting points for the scientometric and bibliometric community, but also to promote the respective on-going activities in Austria.

Our thanks go to the ISSI board for their trust and their constant support, all the contributors for their submissions, the members of the Local and International Committee for their reviewing effort as well as the sponsors for their generous financial support.
We are particulary grateful for the engagement of Heike Faustmann, Alfred Kerschenbauer, Nikolaus Ortner, Johannes Sorz, Silvia Steinbrunner, and Maria-Elisabeth Züger.

Last but not least each conference should also be a feast for all senses. Every endeavour has been made to not only put together an outstanding scientific programme, but also to organize interesting and diverse social events, which will allow you to embrace the beauty and cultural richness of Vienna and its surroundings.

We wish you a great time at the 14th International Society of Scientometrics and Informetrics Conference!

Juan Gorraiz, Edgar Schiebel, Christian Gumpenberger,
Marianne Hörlesberger, and Henk Moed

INDEX

[1] Research in progress paper

THE *CITER-SUCCESS-INDEX*: AN INDICATOR TO SELECT A SUBSET OF ELITE PAPERS, BASED ON CITERS

Fiorenzo Franceschini[1], Domenico Maisano[2] and Luca Mastrogiacomo[3]

[1] *fiorenzo.franceschini@polito.it* [2] *domenico.maisano@polito.it*
[3] *luca.mastrogiacomo@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering), Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract

The goal of this paper is introducing the *citer-success-index* (*cs*-index), i.e., an indicator that uses the number of different citers as a proxy for the impact of a generic set of papers. For each of the articles of interest, it is defined a comparison term – which represents the number of citers that, on average, an article published in a certain period and scientific field is expected to "infect" – to be compared with the actual number of citers of the article. Similarly to the recently proposed *success*-index (Franceschini et al., Scientometrics 92(3):621-6415, 2011), the *cs*-index allows to select a subset of "elite" papers.

The *cs*-index is analyzed from a conceptual and empirical perspective. Special attention is devoted to the study of the link between the number of citers and cited authors relating to articles from different fields, and the possible correlation between the *cs*- and the *success*-index.

Some advantages of the *cs*-index are that (i) it can be applied to multidisciplinary groups of papers, thanks to the field-normalization that it achieves at the level of individual paper and (ii) it is not significantly affected by self citers and recurrent citers. The main drawback is its computational complexity.

Conference Topic

Scientometrics Indicators: Criticism and new developments, Relevance to Science and Technology (Topic 1).

Introduction and Literature Review

In bibliometrics, one of the main analysis dimensions is the impact of scientific publications, which is commonly estimated by counting the number of citations that they accumulate over time (Egghe and Rousseau, 1990). As an alternative to citations, Dieks et al. (1976) and Braun et al. (1985) suggested to use the total number of different citers (or citing authors), i.e., the members of the scientific community who are "infected" by a certain paper. The number of different citers is a proxy which is harder to compute, but more elegant, as only marginally affected by citations from self citers and recurrent citers.

The idea of citers was recently dug up by Ajiferuke and Wolfram (2010), who proposed and implemented an indicator based on citers, without encountering the computational obstacles of the past, thanks to the current evolution of databases and information management tools. The indicator is the *ch*-index, defined for a generic group of papers (e.g., those of a scientist, journal or entire research institution) as *the number (ch) such that, for a general group of papers, ch papers are cited by at least ch different citers while the other papers are cited by no more than ch different citers*. It can be immediately noticed that this definition is similar to that of the *h*-index, with the only exception that, for each publication, the citations obtained are replaced by the number of different citers (Hirsch, 2005).

The *ch*-index was empirically analyzed by Franceschini et al. (2010). This study showed: (i) the general correlation between *ch* and *h*, and (ii) the potential of *ch* in complementing the information given by *h*. E.g., paradoxical situations in which the number of citations obtained by a paper and the number of different citers do not go hand in hand are not so rare, due to the anomalous incidence of recurrent or self citers. A theoretical interpretation of the correlation between *ch* and *h* was recently provided by Egghe (2012).

In this article we focus the attention on the *success*-index (*s*-index), i.e., a recent indicator that, for a generic set of articles, allows to select an "elite" subset, according to a logic different from that of *h* (Franceschini et al., 2012a). The *s*-index is defined as *the number of papers with a number of citations greater than or equal to CT_i, i.e., a generic comparison term associated to the i-th publication.* CT_i is an estimate of the number of citations that articles of the same scientific context and period of time of that of interest (i.e., the *i*-th publication) are likely to achieve.

With the aim of formalizing this definition, a score is associated to each (*i*-th) of the (*P*) publications of interest:

$$\begin{cases} score_i = 1 & \text{when } c_i \geq CT_i \\ score_i = 0 & \text{when } c_i < CT_i \end{cases}, \tag{1}$$

where c_i are the citations obtained by the *i*-th publication. The *s*-index is therefore given by:

$$s\text{-index} = \sum_{i=1}^{P} score_i \ . \tag{2}$$

Apart from *s*, there are other indicators in the literature that allow to select an elite subset, based on the comparison between the number of citations accumulated by each paper and a threshold. E.g., let us consider the selection by $P_{top\ 10\%}$-indicator (Bornmann, 2013), that by π-indicator (Vinkler, 2011), the characteristic scores and scales (CSS) method (Glänzel, 2011) or the ESI's Highly Cited Papers method (ISI Web of Knowledge, 2012). We remark that, differently from *s*, the aforementioned methods require that the set of publications examined are preliminarily categorized into scientific (sub-)disciplines.

As regards the s-index, there are several options for constructing the CT_i related to an i-th paper of interest. Generally, three issues are crucial (Franceschini et al., 2012b):

1. Defining the procedure for selecting a reference sample of homologous publications. Possible approaches are: (i) the selection of papers of same age, type (e.g. research article, review, letter, etc.) and published by the same journal of the i-th paper of interest, (ii) the use of superimposed classifications such as ISI subject categories, (iii) the implementation of "adaptive" techniques in which the sample is determined considering the "neighbourhood" of the paper of interest – typically consisting of the set of papers citing or being cited by it.
2. Deciding whether to consider (i) the distribution of the number of references given or (ii) the citations obtained by the publications of the sample.
3. Identifying a suitable (central tendency) indicator for obtaining CT_i from the distribution of interest, e.g., mean, median, harmonic mean, percentiles, etc..

Regarding point (2), Franceschini et al. (2012a, 2012c) state that indicators based on the distribution of references given – rather than citations obtained – have several advantages:

- The number of references is fixed over time, while the number of citations obtained tends to increase and requires a certain accumulation period to stabilize.
- This stability is also derived by the fact that the number of references is likely to be less variable than the number of citations obtained.
- Bibliographic references are less influenced by journal particularities, such as the average citation impact of articles.

Conceptually, the link between references given (by the papers of the reference sample) and citations obtained (by the papers of interest) originates from a simple consideration: focussing on the totality of the scientific literature in a certain field and according to a simplified model configuration of *isolated* fields – i.e., excluding transfers of citations between different disciplines – the following relationship applies:

$$\sum_{i=1}^{P} c_i = \sum_{i=1}^{P} r_i \ , \tag{3}$$

where

P is the total number of articles (that can cite each other) in the isolated field;
c_i is the number of citations obtained by the i-th paper;
r_i is the number of citations given by the i-th paper.

The equality of Eq. 3 can also be expressed in terms of average values:

$$\frac{1}{P}\sum_{i=1}^{P} c_i = \frac{1}{P}\sum_{i=1}^{P} r_i \quad \Rightarrow \quad \bar{c} = \bar{r} \ . \tag{4}$$

For more detailed and rigorous information on the relation between the $\bar{c}$ and $\bar{r}$ values concerning a set of documents, we refer the reader to (Egghe & Rousseau, 1990).

Returning to the *s*-index, apart from the simplicity of meaning, a great advantage is that it implements a field-normalization at the level of single paper and can therefore be applied to multidisciplinary groups of articles, for instance the whole production output of a research institution.

Another important quality of the *s*-index is that it is defined on a *ratio* scale. This feature has several practical implications that make this indicator more versatile than others – such as the *h*-index, which is defined on an *ordinal* scale (Franceschini et al., 2012a):

- The *s*-index reflects compositions of the input publication sets (with the corresponding citations). In other terms, the union of two groups of publications with *s*-index of 2 and 5 (with no common publications) will always originate a third group of publications with *s*-index of 2 + 5 = 7. This simple property is very useful for extending the use of the *s*-index to multi-disciplinary institutions, e.g., joining groups of publications from different scientific fields.
- The *s*-index eases normalizations aimed at obtaining the so-called size-independency (Franceschini et al., 2012c). Given a general group of papers and the same capacity of producing successful papers, it is reasonable to assume that thr *s*-index should increase proportionally with the different types of "resources" deployed. In fact, several normalized indicators can be obtained dividing the *s*-index by the resource unit of interest; e.g, the staff number of a research institution, the age of a researcher, the number of articles of a journal, the amount of funding received in a certain period, etc..

The purpose of the paper is introducing the *citer-success*-index (or *cs*-index), i.e., a variant of the *s*-index, which is based on citers instead of citations, according to a logic similar to that of *ch*. Given a set of articles, the *cs*-index identifies a subset for which the number of different citers of an *i*-th article exceeds a specified comparison term cCT_i. Formalizing, a score is associated to each *i*-th of the (P) publications of interest:

$$\begin{cases} score_i = 1 & \text{when } \gamma_i \geq cCT_i \\ score_i = 0 & \text{when } \gamma_i < cCT_i \end{cases}, \tag{5}$$

where γ_i are the unique citers related to the *i*-th publication. The word "unique" means that repeated citers are counted only once. The *cs*-index is therefore given by:

$$cs\text{-index} = \sum_{i=1}^{P} score_i \tag{6}$$

Figure 1(a) exemplifies the calculation of the *s*- and *cs*-index for a fictitious set of papers.

In analogy with CT_i, cCT_i is an estimate of the number of unique citers that articles homologous to that of interest are likely to “infect”.

(a)

paper no.	c_i	CT_i	γ_i	cCT_i	s-elite	cs-elite
1	115	20.3	297	60.1	✓	✓
2	86	21.2	187	71.0	✓	✓
3	17	14.5	31	44.8	✓	×
4	15	20.4	68	72.4	×	×
5	12	11.8	30	29.2	✓	✓
6	9	15.7	12	61.9	×	×
					s-index=4	cs-index=3

(b)

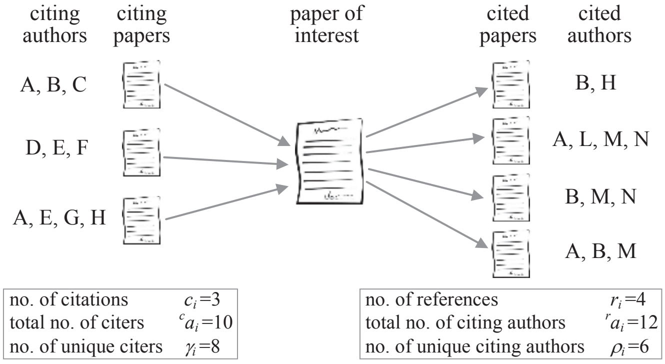

Figure 1. Propaedeutic examples: (a) calculation of the *s*- and *cs*-index for a fictitious set of papers, and (b) introduction of some indicators concerning the authors (represented by letters, e.g., A, B, C, etc.) of papers citing/cited by a fictitious paper of interest.

Similarly to CT_i, there are three basic steps when constructing the cCT_i relating to an i-th article of interest:

1. Selecting a sample of articles homologous to that interest.
2. Deciding whether to consider the distribution of (i) unique citers or (ii) unique cited authors, relating to the papers of the sample.
3. Defining cCT_i by an indicator of central tendency, applied to the distribution chosen at point (2).

The choice at point (2) is more delicate than in the case of the s-index. Intuitively, it may appear convenient to use the distribution of unique cited authors for the same reasons for which, in the case of the s-index, it was convenient to use the distribution of references. However, the link between unique citers and unique cited authors is not necessarily similar to that between r_i and c_i values; even in a model configuration of isolated fields:

$$\sum_{i=1}^{P} \gamma_i \text{ is not necessarily} = \sum_{i=1}^{P} \rho_i , \qquad (7)$$

being

P the total number of papers in the isolated field;

γ_i the number of unique citers of the i-th paper;

ρ_i the number of unique authors cited by the i-th paper.

The reason for this lack of parallelism is twofold and will be examined later in the manuscript.

The rest of the paper is structured in three sections. The section “General link between citers and cited authors” investigates whether it is appropriate to construct the cCT_i by using the distribution of the number of unique authors cited by a sample of papers. The section “Preliminary Empirical analysis of the *cs*-

index" delves into the issue raised in the previous section, examining a large number of papers from different fields. After defining the cCT_i properly, it is studied the correlation between the *s*- and the *cs*-index. Finally, the section "Further remarks" summarizes the original contributions of the paper and the main advantages and disadvantages of the *cs*-index.

General link between citers and cited authors

Before getting into the problem, Figure 1(b) introduces the reader to the indicators and notation that will be used in the remaining of the paper.

Even modelling a scientific field as isolated and considering the totality of the scientific production in it, there are two possible elements of diversity among citing and cited papers: (i) different average number of authors per paper, and (ii) different percentage of unique authors. Let us clarify this point with simple mathematical considerations. The quantity $\sum_{i=1}^{P}\gamma_i$ can be expressed as:

$$\sum_{i=1}^{P}\gamma_i = \left(\sum_{i=1}^{P}\gamma_i \bigg/ \sum_{i=1}^{P}{}^{c}a_i\right)\cdot\left(\sum_{i=1}^{P}{}^{c}a_i \bigg/ \sum_{i=1}^{P}c_i\right)\cdot\sum_{i=1}^{P}c_i = {}^{c}p\cdot{}^{c}app\cdot\sum_{i=1}^{P}c_i, \quad (8)$$

in which

γ_i is the number of unique citers of the *i*-th paper in the isolated field;

${}^{c}a_i$ ($\geq \gamma_i$) is the total number of citers (even repeated, in the case that some citing papers are (co-)authored by the same individuals) related to the *i*-th paper;

c_i is the number of citing papers (or the number of citations obtained) relating to the *i*-th paper;

P is the total number of articles in the isolated field.

As shown in Eq. 8, the quantity $\sum_{i=1}^{P}\gamma_i$ can also be seen as the product of three terms:

${}^{c}p$ $= \sum\gamma_i / \sum{}^{c}a_i$ (≤ 1) i.e., the percentage of unique citers;

${}^{c}app$ $= \sum{}^{c}a_i / \sum c_i$ (≥ 1) i.e., the average number of authors per citing paper;

$\sum_{i=1}^{P}c_i$ the total number of citations obtained.

A "decomposition" similar to that of Eq. 8 may apply to the quantity $\sum_{i=1}^{P}\rho_i$:

$$\sum_{i=1}^{P}\rho_i = \left(\sum_{i=1}^{P}\rho_i \bigg/ \sum_{i=1}^{P}{}^{r}a_i\right)\cdot\left(\sum_{i=1}^{P}{}^{r}a_i \bigg/ \sum_{i=1}^{P}r_i\right)\cdot\sum_{i=1}^{P}r_i = {}^{r}p\cdot{}^{r}app\cdot\sum_{i=1}^{P}r_i, \quad (9)$$

in which

ρ_i is the number of unique authors cited by the *i*-th paper in the isolated field;

$^r a_i$ ($\geq \rho_i$) is the total number of cited authors (even repeated, in the case that some cited papers are (co-)authored by the same individuals) related to the i-th paper;

r_i is the number of papers cited (or the number of bibliographic references) relating to the i-th paper;

P is the total number of articles in the isolated field.

Similarly to $\sum_{i=1}^{P} \gamma_i$, $\sum_{i=1}^{P} \rho_i$ can be seen as the product of three terms:

$^r p = \sum \rho_i / \sum {}^r a_i$ (≤ 1) i.e., the percentage of unique cited authors;

$^c app = \sum {}^r a_i / \sum r_i$ (≥ 1) i.e., the average number of authors per cited paper.

$\sum_{i=1}^{P} r_i$ the total number of references given.

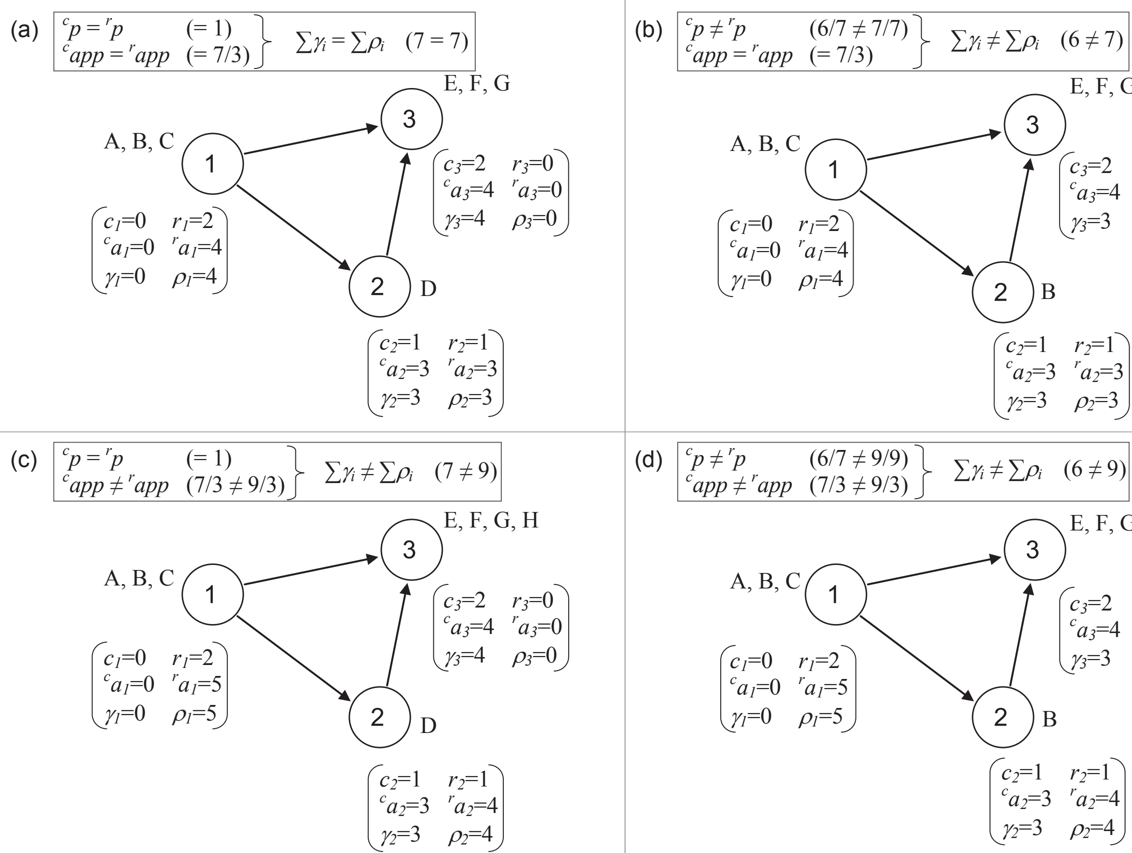

Figure 2. Examples of isolated groups of three papers. Nodes represent the papers (1, 2 and 3), whose authors are A, B, C, D, etc.; arrows represent the citations given by one paper to another. For each paper, it is reported the number of citations obtained (c_i), the number of references given (r_i), the number of total citers ($^c a_i$), the number of total cited authors ($^r a_i$), the number of unique citers (γ_i) and the number of unique cited authors (ρ_i). The equality of Eq. 7 is satisfied in case (a) only, when $^c p = {}^r p$ and $^c app = {}^r app$.

Combining Eqs. 8 and 9 with Eq. 3, it is obtained:

$$\sum_{i=1}^{P}\gamma_i = \left(\frac{^c p}{^r p}\cdot\frac{^c app}{^r app}\right)\cdot\sum_{i=1}^{P}\rho_i \ . \qquad (10)$$

The "balanced" situation $\sum\gamma_i = \sum\rho_i$ can be achieved in the case the following two (sufficient but not necessary) conditions occur (also see the exemplification in Figure 2):

$$\begin{aligned} ^c p &= {}^r p \\ ^c app &= {}^r app \end{aligned} \qquad (11)$$

that is to say, (i) equal average percentage of unique authors and (ii) equal average number of authors for the papers citing and being cited by the total P papers in the isolated field.

Eq. 7 could also be met without necessarily satisfying the two conditions in Eq. 11, that is to say in the case the quantity in brackets in Eq. 10 was unitary. However, there is no practical reason that justify the occurrence of this coincidence, which is purely conjectural. On the other hand, the two conditions of Eq. 11 seem reasonable for (citing and cited) papers within the same field. In any case, they will be tested empirically in the next section.

Table 1. List of journals analyzed within seven ISI subject categories (WoS). For each journal, we considered the research papers issued in the three-year period from 2008 to 2010.

Discipline (ISI Subject Category)	Journal and abbreviation	No. of papers			
		2008	2009	2010	Total
Biology	Bio1 - Bioscience	84	65	66	215
	Bio2 - Biology Direct	46	41	65	152
	Bio3 - Journal of Biosciences	60	65	52	177
Chemistry (analytical)	Che1 - Analytical Sciences	264	238	209	711
	Che2 - Journal of Chemometrics	83	68	76	227
	Che3 - Microchemical Journal	85	114	151	350
Engineering (manufacturing)	Eng1 - International J. of Machine Tools & Manufacture	164	139	118	421
	Eng2 - Robotics and Computer-Integrated Manufacturing	77	96	87	260
	Eng3 - Journal of Intelligent Manufacturing	57	62	71	190
Mathematics	Mat1 - Computational Complexity	20	20	21	61
	Mat2 - Constructive Approximation	31	46	38	115
	Mat3 - Advances in Mathematics	169	146	190	505
Medicine (general & internal)	Med1 - American Journal of Medicine	112	98	119	329
	Med2 - Mayo Clinic Proceedings	86	55	74	215
	Med3 - Medicine	33	40	30	103
Physics (applied)	Phy1 - Applied Physics Express	341	339	345	1025
	Phy2 - Current Applied Physics	177	430	436	1043
	Phy3 - Journal of Magnetic Resonance	230	214	241	685
Psychology	Psy1 - Journal of Experimental Psychology: Learning Memory and Cognition	66	94	52	212
	Psy 2 - Cognitive Psychology	18	26	24	68
	Psy 3 - Health Psychology	125	90	73	288

Preliminary empirical analysis of the *cs*-index

Data collection

A preliminary empirical analysis of the *cs*-index is performed by selecting some papers from a set of journals of seven different ISI subject categories (in brackets the total number of journals indexed by Thomson Scientific in each category): Biology (85), Analytical Chemistry (73), Manufacturing Engineering (37), Mathematics (289), General & Internal Medicine (155), Applied Physics (125), Psychology (75). For each discipline, we selected a random sample of three scientific journals. For each journal, we considered as articles of interest those produced in the three-year period from 2008 to 2010, limiting the selection to research papers only (other document types, such as reviews, conference papers or letters, were excluded). Table 1 contains the journal titles and the number of articles examined for each year. Data are retrieved by querying the Web of Science[1] (WoS) database (Thomson Reuters, 2012).

For each *i*-th article of interest, the following operations are performed.

1. Collection of the citation statistics, consisting of:
   - c_i the number of citing papers published in 2011 and indexed by the database in use;
   - ${}^{c}a_i$ the total number of authors of the (c_i) citing papers (even repeated, if different citing papers are (co-)authored by the same individuals);
   - γ_i the total number of unique citers, obtained by performing the union of the (${}^{c}a_i$) total citers and removing those repeated.

   The choice of a time window for citations accumulation of one year (2011) is to simplify the analysis.
2. Determination of an appropriate cCT_i, which takes into account the propensity to obtain citations from different authors. The construction of cCT_i is based on a sample of *S* articles that are issued in 2011 by the same journal of the (*i*-th) article of interest.
   For each *j*-th of the articles of the sample, we determine:
   - r_j the number of cited papers that were published in the three-year period from 2008 to 2010 and are indexed by the database in use. These constraints were introduced to be consistent with the time window described at point (1) (Moed, 2011);
   - ${}^{r}a_j$ the total number of cited authors (even repeated, if different cited papers are authored by the same individuals);
   - ρ_j the total number of unique cited authors, obtained by the union of the (${}^{r}a_i$) total cited authors, removing those repeated.

   Next, the distribution of the ρ_j values (relating to the papers of the sample) is constructed and the cCT_i is defined by an appropriate central tendency indicator – e.g., the mean ($\overline{\rho}$) or median ($\widetilde{\rho}$). This construction is based on the assumption that, referring to the *i*-th article, the propensity to be cited by different authors is, on average, reasonably close to the propensity to cite different authors, referring to articles issued by the same journal. According to this construction, articles published in the same journal and in the same year will have the same cCT_i value. Probably, a more rigorous way to estimate the cCT_i – but also computationally more expensive – is to use the distribution of the ρ_j values relating to the articles that cite other articles, issued by the article of interest's journal. For further information about this point, please refer to (Franceschini et al., 2012c).

Table 2. Summary of the analysis results. For each of the journals (in Table 1), we report the indicators illustrated in the "Data collection" sub-section. *Overall* indicators are obtained by aggregating the data relating to the three journals examined in each field.

Field	Journ.	^{c}app	^{r}app	^{c}p	^{r}p	P	C	CPP	h	ch	S	R	cCT_i		cs-index		CT_i		s-index	
													$\bar{\rho}$	$\tilde{\rho}$	$(\bar{\rho})$	$(\tilde{\rho})$	$\bar{r}$	$\tilde{r}$	$(\bar{r})$	$(\tilde{r})$
Bio	Bio1	4.6	5.5	0.95	0.91	215	1131	5.3	14	37	76	792	52.3	35.0	25	38	10.4	9.0	30	35
	Bio2	4.9	6.5	0.94	0.86	152	469	3.1	9	26	59	943	89.4	60.0	3	4	16.0	14.0	2	2
	Bio3	5.3	5.9	0.86	0.93	177	274	1.5	7	19	71	382	29.3	18.0	9	20	5.4	4.0	16	17
	overall	4.8	6.0	0.93	0.89	544	1874	3.4	15	45	206	2117	55.0	35.0	31	57	10.3	8.5	37	52
Che	Che1	4.4	4.5	0.89	0.83	711	905	1.3	7	20	191	1076	21.1	17.0	14	30	5.6	5.0	14	14
	Che2	3.9	3.9	0.92	0.86	227	371	1.6	7	17	65	304	15.8	12.0	22	29	4.7	4.0	15	15
	Che3	4.3	4.3	0.92	0.88	350	948	2.7	9	28	185	1274	25.9	22.0	35	50	6.9	5.0	29	51
	overall	4.3	4.3	0.91	0.86	1288	2224	1.7	10	30	441	2654	22.4	17.0	71	128	6.0	5.0	44	78
Eng	Eng1	3.6	3.3	0.86	0.84	421	1148	2.7	9	23	98	392	11.3	9.0	115	142	4.0	3.0	78	126
	Eng2	3.2	3.1	0.93	0.88	260	374	1.4	6	15	101	229	6.2	5.0	74	86	2.3	2.0	57	57
	Eng3	3.0	2.8	0.90	0.93	190	191	1.0	6	10	78	140	4.6	3.0	41	54	1.8	1.0	43	43
	overall	3.4	3.2	0.88	0.87	871	1713	2.0	10	24	277	761	7.6	5.0	261	341	2.7	2.0	266	266
Mat	Mat1	2.2	2.4	0.92	0.86	61	39	0.6	2	6	19	25	2.7	1.0	11	17	1.3	1.0	11	11
	Mat2	2.5	2.1	0.88	0.80	115	178	1.5	4	8	36	87	4.0	3.0	18	26	2.4	1.0	17	31
	Mat3	1.9	2.0	0.88	0.77	687	912	1.3	7	11	290	819	4.3	3.0	113	157	2.8	2.0	126	126
	overall	2.0	2.0	0.88	0.77	863	1129	1.3	7	13	345	931	4.2	3.0	138	190	2.7	2.0	145	145
Med	Med1	5.3	7.5	0.93	0.91	329	533	1.6	6	25	125	946	51.4	36.0	1	7	7.6	6.0	1	4
	Med2	5.3	6.8	0.92	0.89	215	996	4.6	14	37	75	833	66.8	42.0	12	31	11.1	8.0	18	27
	Med3	5.6	7.7	0.92	0.91	103	489	4.7	10	29	48	424	61.8	45.5	7	12	8.8	7.0	17	20
	overall	5.4	7.3	0.92	0.90	647	2018	3.1	15	44	248	2203	58.1	40.0	26	56	8.9	6.0	45	82
Phy	Phy1	5.8	6.1	0.82	0.81	1025	2919	2.8	17	50	418	2483	29.1	24.0	122	147	5.9	5.0	149	149
	Phy2	4.5	4.8	0.89	0.85	1043	1939	1.9	12	34	526	2573	20.1	14.0	99	160	4.9	4.0	111	111
	Phy3	4.4	4.5	0.87	0.79	685	1579	2.3	11	31	243	1671	24.1	19.0	53	80	6.9	6.0	37	37
	overall	5.1	5.2	0.85	0.82	2753	6437	2.3	17	55	1187	6727	24.1	19.0	270	395	5.7	5.0	287	287
Psy	Psy1	2.9	2.7	0.89	0.79	212	545	2.6	10	18	78	596	16.7	15.0	20	23	7.6	7.0	12	12
	Psy2	2.9	2.5	0.88	0.85	68	298	4.4	7	16	17	172	21.3	19.0	10	11	10.1	9.0	5	5
	Psy3	4.3	4.4	0.93	0.89	288	1245	4.3	12	35	90	738	32.4	26.0	43	58	8.2	7.0	32	41
	overall	3.8	3.5	0.92	0.86	568	2088	3.7	15	37	185	1506	24.7	19.0	87	121	8.1	7.0	50	60

The *cs*-index related to the articles of each journal can be calculated using the cCT_i determined at point (2) (according to Eq. 5). The information at point (2) can also be used to determine the average number of authors (^{r}app) and the percentage of unique authors (^{r}p) of the articles cited by the (S) articles of the sample (see Eq. 9). Similarly, the information at point (1) can be used to determine the average number of authors (^{c}app) and the percentage of unique authors (^{c}p) of the articles that cite the (P) articles of interest (see Eq. 8).

The overall ^{c}app, ^{r}app, ^{c}p and ^{r}p values of the seven fields examined can be estimated by aggregating data related to the three journals considered in each discipline.

Information at point (1) can also be used to build other indicators: C (i.e., total number of citations), CPP (i.e., average citations per paper), h, ch and s. As regards the s-index, we will compare the (c_i) citations obtained by each (i-th) paper with a CT_i represented by the mean or median number of references ($\bar{r}_j$ and $\tilde{r}_j$ respectively) that are given by each (j-th) of the articles of the sample.

Conventionally, all indicators are constructed considering the citations obtained in 2011 and the references given to (cited) articles, issued from 2008 to 2010 and indexed by WoS.

Data analysis

Table 2 summarises the results of the empirical analysis. For each journal, the $C = \sum c_i$ total citing papers are those citing each (i-th) of the P papers of interest, and the $R = \sum r_i$ total cited papers are the ones cited by each (j-th) of the S articles of the sample. All statistics were constructed considering the aforementioned time windows and the papers indexed by WoS.

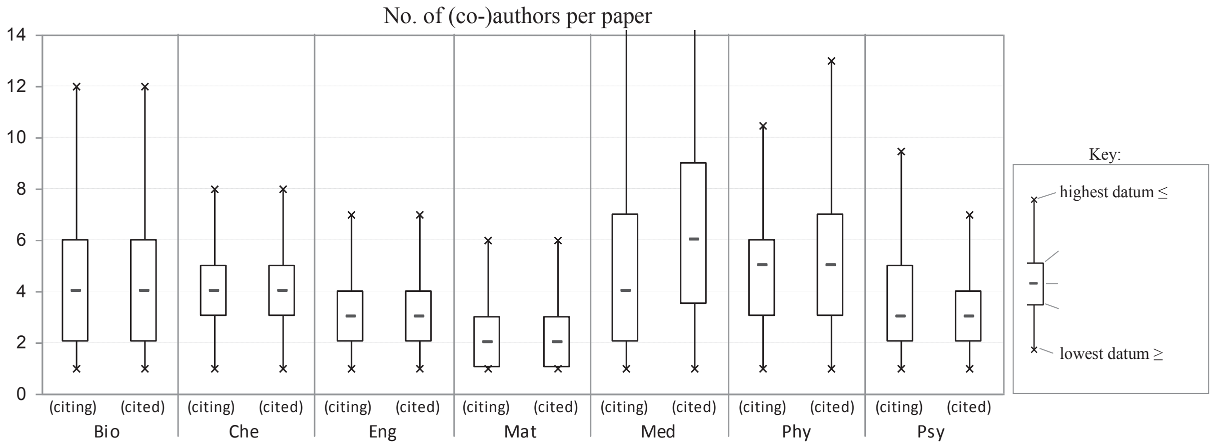

Figure 3. Box-plot of the distribution of the number of (co-)authors relating to the citing and cited papers, concerning the seven fields examined. Citing papers are those that cite the *P* papers of interest while cited papers are those cited by the *S* papers of the macro-sample. $Q^{(1)}$, $Q^{(2)}$ and $Q^{(3)}$ are the first, second and the third quartile of the distributions of interest.

For a specific journal, there are marginal differences between citing and cited authors, as regards (i) the average number of authors per paper (i.e., ${}^{c}app$ and ${}^{r}app$ values) and (ii) the percentage of unique authors (i.e., ${}^{c}p$ and ${}^{r}p$ values).

Besides, there are relatively small variations among the three journals in a specific field. For this reason, it seems appropriate to calculate some aggregated indicators for the whole disciplines (see "overall" indicators in Table 2). The determination of the overall indicators – by joining the data related to the three journals in each discipline – is extended to all the indicators presented in Table 2. In the case of the cs-index and s-index, overall indicators are constructed using cCT_i and CT_i

values determined on the basis of macro-samples obtained by joining the articles issued in 2011 by the three journals selected for each discipline.

Returning to the comparison between ${}^{c}app$ and ${}^{r}app$ values in each field, a simple way to visualize their similarity is through box-plots based on overall statistics. In particular, two distributions are considered; (i) that of the number of authors per paper relating to articles that cite the papers of interest, and (ii) that of the papers cited by the papers of the (macro-)sample (see Figure 3).

It can be seen that, for each discipline, the notches of the two box-plots (respectively for citing and cited papers) almost completely overlap, supporting the view of absence of systematic differences between the two distributions. The same hypothesis can be tested by more rigorous statistical tests, albeit introducing additional assumptions about distributions. On the contrary, when comparing different fields there are systematic differences, confirming what observed in other studies (Glänzel, 2002). For example, let us consider the comparison between the notches relating to Mathematics and Physiscs.

As regards the comparison between ${}^{c}p$ and ${}^{r}p$ values, the question is a bit more complicated: the overall percentages of different authors (respectively citing or cited) can be seen as weighted averages of the same percentages, at the level of individual papers:

$$\begin{aligned} {}^{c}p &= \left(\sum_{i=1}^{P}\gamma_i\right)\Bigg/\left(\sum_{i=1}^{P}{}^{c}a_i\right) = \left(\sum_{i=1}^{P}{}^{c}p_i\cdot{}^{c}a_i\right)\Bigg/\left(\sum_{i=1}^{P}{}^{c}a_i\right) \\ {}^{r}p &= \left(\sum_{j=1}^{S}\rho_j\right)\Bigg/\left(\sum_{j=1}^{S}{}^{c}a_j\right) = \left(\sum_{j=1}^{S}{}^{r}p_j\cdot{}^{r}a_j\right)\Bigg/\left(\sum_{j=1}^{S}{}^{r}a_j\right) \end{aligned}, \qquad (13)$$

being

${}^{c}p_i$ the percentage of unique citers relating to the *i*-th of the *P* papers of interest;

${}^{c}a_i$ the “weight” of ${}^{c}p_i$, i.e., the number of authors (even repeated) citing the *i*-th paper;

${}^{r}p_j$ the percentage of unique authors cited by the *j*-th of the *S* papers of the sample;

${}^{r}a_j$ the “weight” of ${}^{r}p_j$, i.e., the number of authors (even repeated) cited by the *j*-th paper.

Being ${}^{c}p$ and ${}^{r}p$ weighted quantities, one can represent the distributions of ${}^{c}p_i$ and ${}^{r}p_j$ values by special box-plots based on *weighted quartiles*, defined as:

- ${}^{c}Q_w^{(1)}$, ${}^{c}Q_w^{(2)}$ and ${}^{c}Q_w^{(3)}$, i.e., the weighted first, second (or weighted median) and third quartile of the ${}^{c}p_i$ values. These indicators are obtained by ordering in ascending order the ${}^{c}p_i$ values of the articles of interest and considering the values for which the cumulative of weights is equal to respectively the 25%, 50% and 75% of their sum;
- ${}^{r}Q_w^{(1)}$, ${}^{r}Q_w^{(2)}$ and ${}^{r}Q_w^{(3)}$, i.e., the weighted first, second (i.e., the weighted median) and third quartile of the ${}^{r}p_i$ values.

The box-plots relating to weighted quartiles are represented in Figure 4. The differences between the ${}^{c}p_i$ and ${}^{r}p_j$ distributions within the same field seem insignificant. We also note the absence of significant differences between fields.

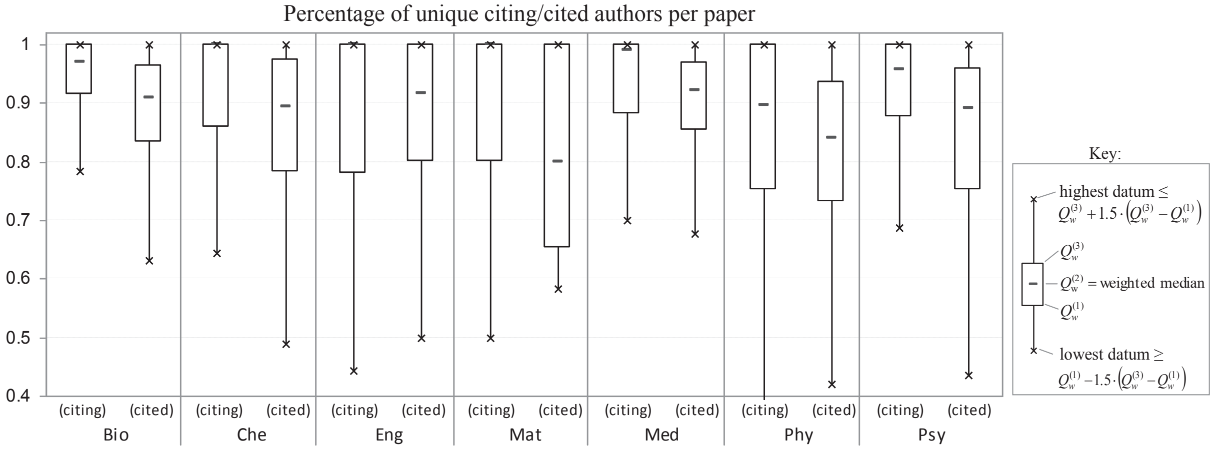

Figure 4. "Weighted" box-plot of the percentage of unique citing (${}^{c}p_i$) and cited authors (${}^{r}p_i$), relating to the papers that cite the papers of interest and are cited by the papers of the macro-sample, in the seven fields examined. $Q_w^{(1)}$, $Q_w^{(2)}$ and $Q_w^{(3)}$ are the first, second and the third weighted quartile of the distributions of interest

Returning to Table 2, there are relatively little differences in terms of cCT_i values (i.e., estimators of the propensity to cite different authors), for journals of the same field. Some exceptions are: Bio2 for Biology and Eng1 for Engineering. This incomplete uniformity is probably due to the fact that some journals are influenced by publications of neighbouring fields, with different citation propensity. For a more rigorous estimate, it would probably be appropriate to define cCT_is using a larger sample of papers/journals.

For each journal, in Table 2 are reported two different cCT_js: i.e., using $\overline{\rho}$ and $\widetilde{\rho}$. In general, the resulting values are higher in the first case. This probably depends on the incidence of papers characterized by hyperauthorship – i.e., literally tens or even hundreds of authors (Cronin, 2001) – which tends to "inflate" $\overline{\rho}$ but not $\widetilde{\rho}$, as the latter indicator is only marginally sensitive to the right tail of the distribution of ρ_j values.

Another interesting aspect is the link between *cs*-index and *s*-index. The diagram in Figure 5 – which is constructed using $cCT_i = \overline{\rho}$ and $CT_i = \overline{r}$ (in Table 2) – shows a strong correlation ($R^2 \approx 89\%$), similar to that between *ch* and *h* (Franceschini et al., 2010; Egghe, 2012). All the points of the graph – although resulting from articles of different scientific fields – tend to be distributed around the same trend line, which is very close to the bisector of the *cs*-*s* plane.

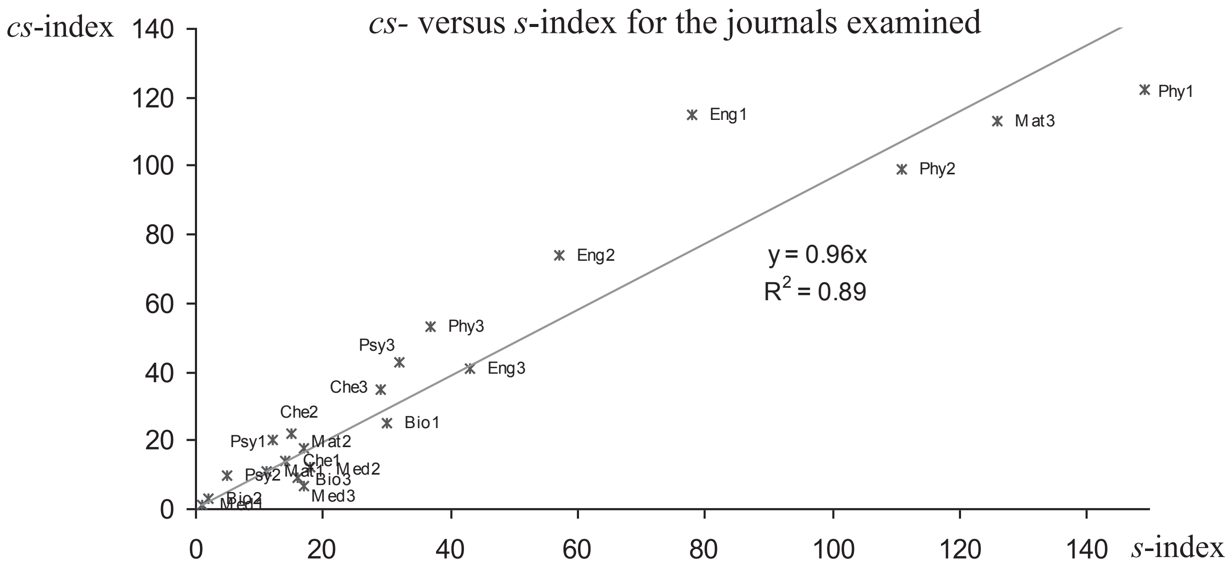

Figure 5. Relationship between the *cs*- and *s*-index for the journals examined. Indicators are calculated considering respectively $cCT_i = \overline{\rho}$ and $CT_i = \overline{r}$ (see Table 2).

In the absence of "anomalies" – e.g., high incidence of self-citations or citations from recurrent citing authors – the *cs*-index and *s*-index should be very close. Therefore, the study of their difference can be useful to highlight abnormal situations. For example, consider the point related to Med3 in Figure 5, which corresponds to a relatively high value of *s*-index, associated to a quite small value of *cs*-index, probably due to a relatively high incidence of self citers and recurrent citing authors. On the contrary, the point related to Eng1 denotes an opposite situation, in which *cs*-index is much larger than *s*-index. probably due to an opposite attitude.

Further remarks

This study revealed some interesting points that it is worth summarizing and developing in the following:

- The analysis suggests that the comparison term (cCT_i) of the *cs*-index can be constructed using the distribution of the ρ_j values related to the papers of a sample. This is justified by the absence of systematic differences between (i) the average number of authors and (ii) the average percentage of unique authors, between citing and cited papers in a certain field. On the other hand, the analysis confirmed some systematic differences between fields, as regards the average number of authors per paper.
- The *cs*-index is an indicator that, although generally correlated with the *s*-index, can complement it, being only marginally affected by self-citations and citations from recurrent citers.
- Similarly to the *s*-index, the *cs*-index has an immediate meaning and is practical for normalizations aimed at obtaining the so-called size-

independency, thanks to the ratio scale property (Franceschini et al., 2012a). For example, scientific journals with a different number (*P*) of articles could be easily compared by means of the percentage of "successful" papers, i.e., *cs*-index/*P*.

- Even if it was not shown directly in this paper, another advantage "inherited" by the *s*-index is that cs-index can be calculated for a set of multidisciplinary articles, thanks to the field-normalization that it achieves at the level of individual paper. For example, the *cs*-index can be used as a proxy for synthesizing the productivity and impact of (i) the whole publication output of scientists involved in multiple disciplines (e.g., mathematicians or computer scientists actively involved in bibliometrics), or (ii) that of entire multidisciplinary research institutions.
- A disadvantage of the *cs*-index is the computational complexity of the cCT_i values. E.g., our data collection and analysis – which was performed by an *ad hoc* application software able to query the WoS database automatically – took about twenty consecutive hours.
- Another potential drawback of *cs*-index is represented by hyperauthorship, which could lead to inflate cCT_i values. A partial solution to this problem is (i) to determine cCT_i by indicators that are insensitive to the right-hand tail of the distribution of ρ_j (e.g., $\tilde{\rho}$), or (ii) to apply some exclusion criteria, so as to curtail the count of the authors of a certain paper, according to a conventional threshold.

[1] The WoS database configuration included the following resources: Citation Index Expanded (SCI-EXPANDED) from 1970 to present, Social Sciences Citation Index (SSCI) from 1970 to present, Arts & Humanities Citation Index (A&HCI) from 1975 to present, Conference Proceedings Citation Index - Science (CPCI-S) from 1990 to present, Conference Proceedings Citation Index - Social Science & Humanities (CPCI-SSH) from 1990 to present.

References

Ajiferuke, I. & Wolfram, D. (2010). Citer analysis as a measure of research impact: Library and information science as a case study. *Scientometrics*, 83(3), 623–638.

Bornmann, L. (2013). A better alternative to the h index. *Journal of Informetrics*, 7(1), 100.

Braun, T., Glänzel, W. & Schubert, A. (1985). Scientometric Indicators: A 32-country comparative evaluation of publishing performance and citation impact. Philadelphia: World Scientific.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, 52(7), 558–569.

Dieks, D. & Chang, K.H. (1976). Differences in impact of scientific publications: Some indices derived from a citation analysis. *Social Studies of Science*, 6(2), 247–267.

Egghe, L. & Rousseau, R. (1990). Introduction to Informetrics: Quantitative Methods in Library, documentation and Information Science, Amsterdam: Elsevier.

Egghe, L. (2012). A rationale for the relation between the citer h-index and the classical h-index of a researcher, to appear in *Scientometrics*, DOI 10.1007/s11192-012-0770-1.

Franceschini, F., Maisano, D., Perotti A. & Proto A. (2010). Analysis of the *ch*-index: an indicator to evaluate the diffusion of scientific research output by citers, *Scientometrics*, 85(1), 203–217.

Franceschini, F., Galetto, M., Maisano, D. & Mastrogiacomo, L. (2012a). The success-index: an alternative approach to the h-index for evaluating an individual's research output, *Scientometrics*, 92(3), 621-641.

Franceschini, F., Galetto, M., Maisano, D. & Mastrogiacomo, L. (2012b). Further clarifications about the success-index, *Journal of Informetrics*, 6(4): 669–673.

Franceschini, F., Maisano, D. & Mastrogiacomo, L. (2012c). Evaluating research institutions: the potential of the success-index, to appear in *Scientometrics*, DOI 10.1007/s11192-012-0887-2.

Glänzel, W. (2002). Coauthorship patterns and trends in the sciences (1980-1998): A bibliometric study with implications for database indexing and search strategies, *Library Trends*, 50(3), 461–473.

Glänzel, W. (2011). The application of characteristic scores and scales to the evaluation and ranking of scientific journals, *Journal of Information Science*, 37(1), 40–48.

Hirsch, J.E. (2005). An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569–16572.

ISI Web of Knowledge (2012). Essential Science Indicators. http://thomsonreuters.com [Nov. 2012]

Moed, H.F. (2011). The Source-Normalized Impact per Paper (SNIP) is a valid and sophisticated indicator of journal citation impact. *Journal of the American Society for Information Science and Technology*, 62(1), 211–213.

Thomson Reuters (2012) 2011 Journal Citation Reports – Science Edition, www.isiknowledge.com.

Vinkler, P. (2011). Application of the distribution of citations among publications in scientometric evaluations, *Journal of the American Society for Information Science and Technology*, 62(10), 1963–1978.