A Model for Publication and Citation Statistics of Individual Authors

Wolfgang Glänzel[1, 2], Sarah Heeffer[1], and Bart Thijs[1]

[1] *{wolfgang.glanzel, sarah.heeffer, bart.thijs}@kuleuven.be*
[1]KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
[2] Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract

One of the most important requirements of building applicable models and meaningful indicators for the use of scientometrics at the micro and meso level is the correct identification and disambiguation of authors and institutes. Platforms like ResearcherID or ORCID with author registration providing high reliability but lower coverage now provide appropriate data sets for the development and testing of stochastic models describing the publication activity and citation impact of individual authors. This paper proposes a triangular model incorporating papers, citations and authors analogously to the dichotomous model used at higher levels of aggregation like countries or fields. This model is applied to a set of authors in any field of science identified by their ResearcherID. However, the main advantage of classical citation indicators to study citation impact under conditional productivity turned out to be the main problem in this triangle: the possible heterogeneity of the collaborating authors results in low robustness. A mere technical solution to this problem would be fractional counting at three levels, but the conceptual issue, the different roles of co-authors causing this heterogeneity, will never be solved by any algorithm.

Conference Topics

Methods and techniques; Data Accuracy and disambiguation

Introduction

Spectacular progress has been made in author identification, the disambiguation of names and their institutional assignment on the basis of correct affiliation and cleaned address data extracted from bibliographic databases. In particular, this is one of the most important and basic requirements of building applicable models and meaningful indicators for the use of scientometrics at the micro and meso level. Correct author identification is not only indispensable in studies of academic careers, researchers' mobility, authors' publication and collaboration patterns (Braun et al., 2001) but also in monitoring constitution and performance of research teams (Strotman & Zhao, 2012). The task outlined here is practically twofold: On the one hand, the large-scale disambiguation and assignment of authors forms still one of the big challenges in scientometrics. Although the quality of disambiguation and assignments of authors has considerably improved due to sophisticated algorithms and scientometric techniques, e.g., using "bibliometric fingerprints" (Tang & Walsh, 2010) and similarity patterns (cf. Caron & van Eck, 2014), automated processes proved not sufficient to provide reliable reference standards even if optional interaction of individual authors has been made possible. In this context author identification of the Mathematical Reviews and Elsevier's Scopus databases might just serve as examples. Mathematical Reviews was one of the first databases that applied automated processes (since 1985) for author identification. Challenges are, among others, mobility, topic shifts, career breaks, occasional and infrequent publication activity, e.g., so-called transients (Price & Gürsey, 1976). Incorrect institutional assignment, multiple identities as well as unresolved homonyms are still frequently observed errors. This is contrasted by the possibly higher reliability but lower coverage of identifiers that are based on author registration as, for instance, the ResearcherID of the Web of Science database (Thomson Reuters) and the Open Researcher and Contributor ID (ORCID). The latter IDs are sensitive to human errors and their willingness to regularly update and maintain publication assignment to their IDs. A previous study has pointed to the representativeness bias in favour of more prolific authors (Heeffer et al., 2013).

The second issue is partially related to methodology but also of conceptual nature. The methodological issues arise from the superposition of multiple assignments of publication to subjects, on the one hand, and to co-authors and their particular profiles, on the other hand.

Stochastic models for publication activity and citation impact of authors, however, require partitions, which can only partially approximated by corresponding fractionation procedures (cf. Glänzel et al., 2014 for multiple subject assignment in the context of Characteristic Scores at Scales at different levels of aggregation). A further issue arises from the different stages of the individual careers of authors at the same time; while the same publication year ensures the same age of papers in a given citation window, a pre-set publication year collects papers of scientists who are situated in completely different stages of their careers at the same time. The fact that a PhD student or post-doctoral fellow might collaborate with a senior scientist makes the situation even more complex. Thus the question arises whether the same reference standard derived from the data set should apply to the junior as to the senior co-author. And this leads us directly to the conceptual problem: What is the weight of co-authors and their profiles in determining standards for possible benchmarking exercises? This implies that large-scale statistics calculated on the basis of given publication periods and selected subject fields will not be appropriate as reference standards at the micro level but might indeed mirror the profiles of larger institutions and countries adequately and thus serve as general model at these levels of aggregation.

In this paper a triangular stochastic model analogously to the models used at higher levels of aggregation will be described and opportunities and limitations of such a model will be discussed. In the following we will mainly focus on the following questions.

1. What is the relationship between authors' productivity and their citation impact?
2. How can the relationship between the authors' citation impact and the impact of their publications be described?
3. What is the possible effect of co-authorship on these patterns?
4. Can any reference standard for evaluative studies be derived from the model and the empirical data?

This short introduction already adumbrates the possibilities but also the limitations of scientometric models that are created on the basis of the identification and assignment of individual authors. We optionally attempted to use Thomson Reuters' *Distinct Author Identification System* (DAIS), which is based on clustering author names, institution names, and citing and cited author relationships (Thomson Reuters, 2012). As all automated processes, this results in a broader coverage, but suffers from false positives. We have found nearly 30 authors with more than 300 papers each in 2011 according to the DAIS and the most productive author had 1272 WoS indexed papers. However, a simple manual check of names and profiles of authors associated with the same DAIS code revealed different persons with the same family name and first initial but partially different given names and different research profiles. In order to reduce uncertainty we decided therefore to use Thomson Reuters' ResearcherID in conjunction with journal articles published in the same year hazarding the consequences of representativeness bias. From the viewpoint of the model and the analysis this restriction is, however, immaterial. In this context we would like to stress again that the possible biases in representativeness of author selection is insignificant from the viewpoint of the creation and applicability of the model. More important in this context is the reliability of identification of the authors and their affiliations. Nevertheless, we will first have a look at representativeness of author selection on the basis of Thomson Reuters' ResearcherID (RID). This first part of the analysis forms a straight continuation of a previous study on productivity of registered authors by Heeffer et al. (2013).

Data sources and data processing

All papers indexed as articles, proceedings papers, reviews and letters in the 2011 volume of Thomson Reuters Science Citation Index Expanded (SCIE) have been selected. The reason for this choice of a single year publication window, which results from structural properties of author representation and productivity reflected by annual document indexation in bibliographic databases, is as follows. We have already mentioned at the outset that citation processes of scientific papers published in the same year have the following properties: Within a given citation window, all documents in the set have the same age at any particular time and the citation process is not *homogeneous*, that is, citation frequencies at the initial period differ from those at later stages. Paradigmatically this phenomenon has been characterised as a combination of phases of maturing and decline in citation processes (Glänzel & Schoepflin, 1995; Moed et al., 1998). As a consequence, enlarging the citation window will not simply result in a multiplication of citations by a factor proportional to the length of the window. The situation is completely different when a population with heterogeneous age structure is underlying the process and authors are constantly entering and leaving the system. While the citation process of a fixed document set can be described, for instance, by a simple birth process (e.g., Glänzel & Schoepflin, 1994), the publication distribution of an author set, which is subject to changes and interacts with the "environment", requires a different model taking also the effect of immigration and emigration into account. Such model has been proposed by Schubert and Glänzel (1984). This is the situation we find in any publication period in a bibliographic database: Newcomers are entering the author population, terminators are leaving the system and continuants are members of the population for a longer time including the complete period under study (cf. Price & Gürsey, 1976). As a consequence, publication activity in a longer time period can be simulated by multiplying productivity by a proportionality factor according to the length of the period. Therefore it is initially sufficient to select a shorter period of, e.g., one year as the basis of the analysis.

The reason why we have chosen the year 2011 was that in this particular year the share of papers with registered RID was the largest. We expected, of course, that this share will increase and that more authors will be registered in more recent years but the fact that this share decreases beyond 2011 is probably caused by the attitude of authors to update registration and register newly indexed papers not always immediately and regularly but rather intermittently. The choice of 2011 was also convenient because it allows the observation of citations in an appropriate time span. In addition to this publication year we could therefore choose the three-year citation window 2011-2013.

Methods and results

Theoretical considerations

As already mentioned in the previous section, the inclusion of productivity patterns in citation statistics permits insight into a complex system with the provision of a whole set of benchmarks and reference values. From the mathematical viewpoint, we deal with two basic variables that can stochastically be considered random variables, ζ expressing publication activity and ξ standing for citation rates. Yet the two variables are not assumed to be independent and it is commonly known that more prolific authors tend to be more cited as well. Therefore $P(\xi=i|\zeta=j)$ does not necessarily equal $P(\xi=i)$ for all $i, j \geq 0$ and the conditional expectation $E(\xi|\zeta=j)$, being a function of ζ and taking its values with probability $P(\zeta=j)$ is not necessarily constant. In our case, the following measurable variables occur: The publication activity of a (randomly chosen) author in the mirror of the SCIE database in 2011, the citation impact of a (randomly chosen) paper indexed in the 2011 volume of the SCIE and the citation impact of an author with one or more papers in 2011 with the intermediate conditional measure of citation impact, provided the author has a given number of publications $j \geq 0$ in 2011.

The following mathematical description, which is indeed necessary to avoid confusions, will, however, be restricted to the absolute necessary. The first question formulated in the introduction relates to the relationship between authors' productivity and their citation impact. This can be formulated as follows. Since citation impact is always measured through the citation rates of individual publications, an author's citation impact can theoretically be obtained as

$$P(\xi=i) = \Sigma_j\, P(\xi=i|\zeta=j)\cdot P(\zeta=j) \text{ for all } i \geq 0,$$

with the corresponding expectation

$$E(\xi) = \Sigma_j\, E(\xi|\zeta=j)\cdot P(\zeta=j).$$

Index j is assumed to be positive because the trivial case $P(\xi=i|\zeta=0) = 1$, if $i = 0$ and $P(\xi=i|\zeta=0) = 0$, otherwise, can be excluded (no citations without publications). The corresponding statistics are then denoted as $f_i|j$ and $x|j$. Both statistics (conditional empirical distribution and mean value) refer to the citation impact of authors. Furthermore, the corresponding conditional mean citation rate of an author's *papers* can be obtained by dividing $x|j$ by the number of papers j, that is, $(x|j)/j$ with $j > 0$ is an estimator of the expected citation rate of the individual papers of an author with j papers in the given publication year.

In order to tackle the second problem, we have to introduce a third variable, which will complete the triangular model. Using the notation η for the citation impact of a single paper by an individual author, we obtain a more complex formula than above for the conditional probabilities taking all possible combinatorial combinations concerning number of publications and their citations into account but the relationship of their expectations simply reduced to $E(\xi) = E(\eta)\cdot E(\zeta)$. Under the simple assumption that the likelihood not to be cited is the same for all papers of the author, i.e., $q = P(\eta=0)$ for all $j > 0$, we can approximate the probability of author uncitedness and citedness as $P(\xi=0) = \Sigma_j\, q^j\cdot P(\zeta=j) = P(\eta=0)^j$ and $P(\xi>0) = 1-P(\xi=0)$, respectively. The reason for the relative simplicity of this expression is that uncitedness of an author in a given period implies that none of his/her papers is cited. The extreme cases $P(\xi=0) = 0$ and $P(\xi=0) = 1$ are obviously equivalent with $q = 0$ and $q = 1$, respectively. We will denote the empirical value of q by g_0. Using the mean values x, z and y as estimators of expected citation rate of an author, the expected publication activity of an author and the expected impact of the author's papers, respectively, we obtain the simple relationship $x = y\cdot z$. From the elementary considerations we can conclude that at least basic statistics can be readily expressed with the aid of two variables.

Finally, it might be worth mentioning in this context that the above random variables and the corresponding statistics also form the groundwork for modelling Hirsch-type indices, notably their cumulative versions such as the successive h-index (e.g., Schubert, 2007).

The sample

The sample of RID authors does – as already observed by Heeffer et al. (2013) – not form a *random sample* of the complete author population in the database as RID authors are less frequent at the low end (particularly among single-paper authors), and are more productive at the high end of the productivity distribution.

Table 1. Share of papers with RID authors and their relative citation impact by countries [Data sourced from Thomson Reuters Web of Science Core Collection].

Country	*Papers*	*RCR*	*NMCR*	*%HC*	*RCR*	*NMCR*	*%HC*	*%RID*
Argentina	7702	1.03	0.98	1.3%	1.44	1.91	4.5%	14.7%
Australia	40979	1.16	1.36	2.1%	1.22	1.63	2.9%	42.4%
Austria	12274	1.22	1.45	2.6%	1.39	2.01	4.7%	29.7%
Belgium	17598	1.22	1.51	2.5%	1.34	1.96	4.1%	32.6%
Brazil	33940	0.99	0.72	0.7%	1.02	0.88	1.0%	45.2%
Canada	54511	1.14	1.38	2.1%	1.38	2.08	4.3%	21.0%
Chile	5073	1.15	1.08	1.3%	1.31	1.49	2.6%	31.8%
Czech Rep.	9350	1.18	1.09	1.5%	1.27	1.40	2.4%	40.4%
Denmark	12772	1.30	1.62	3.1%	1.41	2.03	4.4%	36.2%
Egypt	6251	1.02	0.75	0.6%	1.41	1.52	2.9%	15.1%
Finland	9945	1.20	1.42	2.2%	1.35	1.91	3.8%	34.7%
France	65238	1.09	1.29	1.8%	1.20	1.71	3.0%	28.4%
Germany	91263	1.14	1.39	2.1%	1.23	1.81	3.4%	30.7%
Greece	10647	1.13	1.12	1.6%	1.45	1.91	4.1%	22.2%
Hungary	5763	1.15	1.16	1.8%	1.36	1.63	3.4%	36.2%
India	46532	0.98	0.68	0.7%	1.20	1.26	1.8%	13.0%
Iran	20234	1.15	0.71	0.8%	1.55	1.36	2.8%	9.1%
Ireland	6833	1.18	1.42	2.3%	1.34	1.85	3.5%	35.5%
Israel	11558	1.06	1.34	2.1%	1.28	1.97	4.2%	21.4%
Italy	53919	1.10	1.22	1.7%	1.19	1.52	2.6%	32.8%
Japan	76799	0.94	0.96	1.1%	1.13	1.52	2.5%	20.9%
Malaysia	7325	1.12	0.71	0.7%	1.15	0.84	0.9%	41.1%
Mexico	9830	1.02	0.89	1.2%	1.40	1.69	3.4%	21.0%
Netherlands	31883	1.21	1.60	2.8%	1.28	1.90	3.8%	36.8%
New Zealand	7186	1.17	1.33	2.1%	1.45	1.98	4.0%	30.5%
Norway	9694	1.23	1.43	2.4%	1.43	2.07	4.8%	26.7%
Pakistan	5371	1.18	0.69	1.1%	1.52	1.58	3.3%	16.0%
China PR	156403	1.04	0.91	1.1%	1.24	1.53	2.9%	20.2%
Poland	20261	1.08	0.82	0.9%	1.30	1.41	2.3%	20.1%
Portugal	9844	1.14	1.19	1.6%	1.17	1.29	1.9%	63.9%
Romania	6618	1.26	0.71	1.2%	1.30	0.97	1.9%	40.0%
Russia	27853	1.03	0.55	0.7%	1.12	0.94	1.5%	26.5%
Saudi Arabia	5417	1.15	0.92	1.3%	1.35	1.42	2.4%	31.4%
Singapore	9458	1.17	1.53	2.8%	1.29	1.91	4.1%	47.0%
South Africa	7787	1.26	1.19	2.2%	1.50	1.73	4.4%	25.3%
South Korea	44228	0.97	0.89	1.0%	1.13	1.44	2.4%	22.2%
Spain	47885	1.10	1.24	1.7%	1.19	1.56	2.6%	35.9%
Sweden	19923	1.18	1.44	2.4%	1.31	1.90	3.8%	30.8%
Switzerland	23582	1.29	1.73	3.3%	1.38	2.16	5.0%	34.6%
Taiwan	25550	0.92	0.93	1.1%	1.19	1.55	2.9%	17.0%
Thailand	5819	1.08	0.89	1.0%	1.32	1.48	2.6%	16.9%
Turkey	22571	1.02	0.63	0.8%	1.39	1.34	2.7%	12.8%
UK	91438	1.16	1.46	2.4%	1.28	1.90	3.9%	31.6%
USA	333610	1.09	1.40	2.2%	1.25	1.95	3.9%	20.0%
World total	1229248	1.00	1.00	1.2%	1.13	1.42	2.2%	21.1%

Nevertheless, from the viewpoint of the objectives of this study, this bias is primarily insignificant. In total we have 1,229,248 documents among which 259,341, that is, 21.1% had at least one registered (RID) author. This share considerably varies among countries. The share ranges between about 10% in Africa, Arabic countries and India till about 50% and even more in Brazil, Singapore and Portugal.

Table 1 displays statistics of countries with at least 5,000 publications in 2011. In particular, the variable RCR represents the relation of observed citation impact and the corresponding journal-based expectation, NMCR stands for corresponding relation between observation and discipline-based expectation and %HC is the share of highly cited papers, that is, of papers that have received at least seven times as many citations as the standard of their discipline (see Glänzel et al., 2009 for exact definitions). The last variable %RID, finally, expresses the share of papers with (at least one) author with registered RID. The comparison of relative citation rates and the share of highly cited papers provides empirical evidence that papers by registered authors exhibit distinctly higher citation impact than the corresponding national standards. We would also like to mention that only very few exceptions have been found in smaller countries not displayed here, e.g., Jordan and Latvia, where the share of highly cited papers and the RCR values did not reach their national standards created by all authors.

Representativeness of publications by authors with RID in individual subject fields is in line with our intuitive expectations: The share of papers by RID authors is the lowest in Mathematics (13.0%), clinical and experimental medicine (14.2% for general & internal medicine and 14.2% for non-internal specialties) and engineering (18.7%). This is contrasted by the corresponding shares in physics, chemistry and biosciences (29.8%, 28.5% and 25.0%, respectively).

Productivity and impact of RID authors

The bias in publication-activity statistics of registered authors has already been stressed (cf. Figure 3 in Heeffer et al, 2013). In particular, RID authors are less frequent at the low end, and more productive at the high end of the productivity scale. Figure 1 shows the distribution of papers over RID authors in 2011. The underlying data are based on the short period of only one year so that the share of single-paper authors is consequently large. Nevertheless, the productivity distribution has the expected long tail: 87 authors have (co-)authored more than 50 papers each. We just mention in passing that the maximum count amounted to 296. This almost incredibly large annual publication output of publishing almost one paper a day is, however, formally correct. The author with an affiliation at the University Sains in Malaysia and a second, more recent one at the King Saud University in Saudi Arabia is active in crystallography. In this context we have to notice that the number of his co-authors per paper is rather low, so that even fractionation would not essentially decrease this author's publication count. This example also illustrates that conceptual issues might have more weight than the number or seniority of co-authors. Before we discuss field-specific aspects of authorship statistics, we still have a look at general citation patterns.

In Figure 2, the citation distribution over authors is compared with the corresponding distribution by papers. In addition to the two series of bars expressing the frequency of citations by RID authors and their papers, respectively, a solid line displays the citation distribution of all papers indexed in the SCIE database to illustrate the bias of the sample. The more moderate skewness and greater expectation of the distribution of citations over authors are plausible and in line with the theoretical rudiments described in the previous subsection since usually we have $z > 1$ and $g_0 \in (0, 1)$.

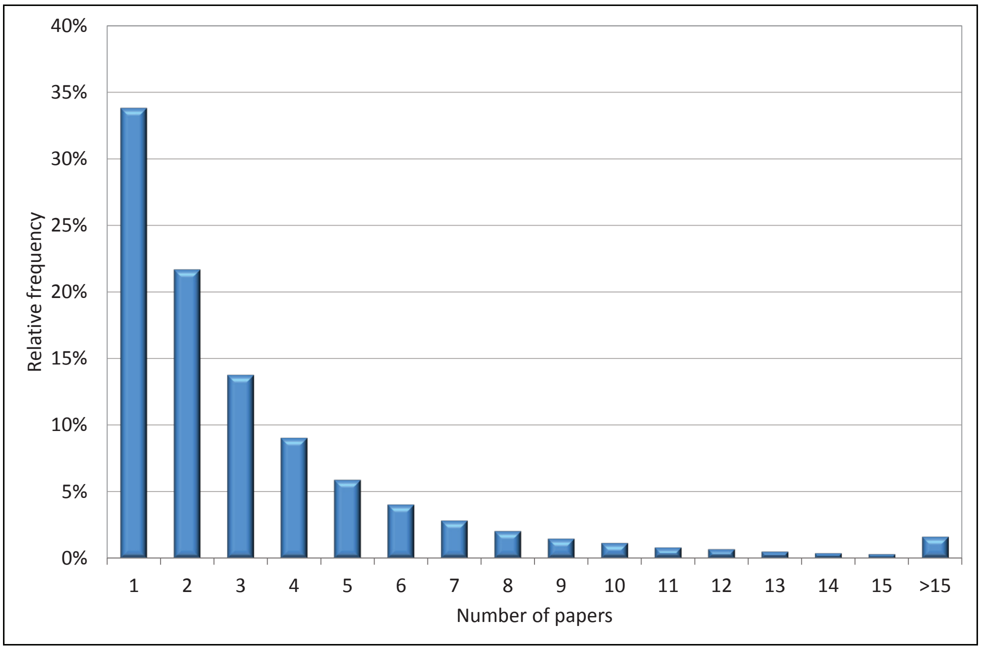

Figure 1. Relative frequency of publication activity of RID authors in 2011. [Data sourced from Thomson Reuters Web of Science Core Collection].

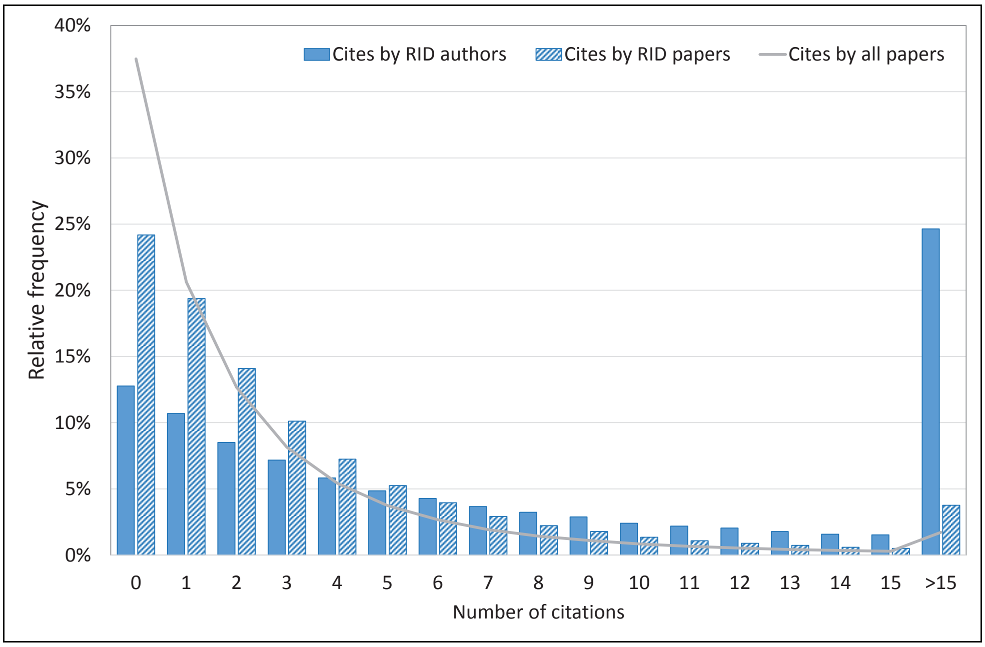

Figure 2. Empirical citation distribution related to RID authors in 2011 in a 3-year citation window. [Data sourced from Thomson Reuters Web of Science Core Collection].

A simple regression analysis aims at studying the relationship of productivity and citation impact of authors, on the one hand, and his/her publications, on the other hand. Conditional mean citation rates in the citation window 2011–2013 received by papers published in 2011 by registered authors have been plotted against their productivity (see Figure 3). Productivity higher than 32 papers has been omitted because of low frequency and considerably fluctuations beyond this level. A power-law model for author citations reflects a very strong correlation, whereas the regression for article citations by authors proved to be linear with somewhat weaker correlation.

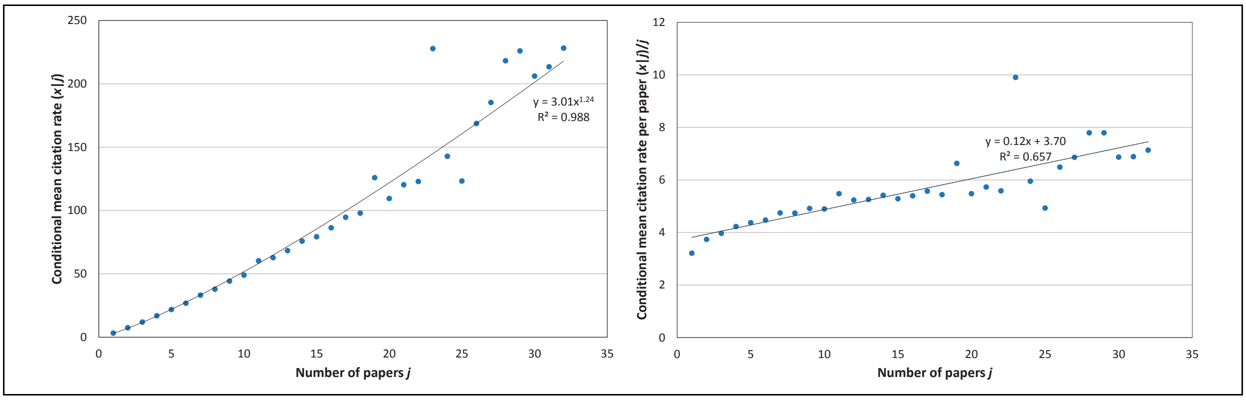

Figure 3. Plot of conditional citation impact of RID authors (left-hand side) and RID papers (right-hand side) based on a 3-year citation window vs. productivity in 2011 [Data sourced from Thomson Reuters Web of Science Core Collection].

While a positive effect of productivity on the expected citation impact of authors was, of course, expected (an increase of papers cannot result in less citations), the positive correlation between number of papers and the mean citation rate of *those* papers is as such not necessarily an inherent property of the model and as we described in the first subsection, the three variables ξ, ζ and η are not assumed to be independent. This indeed substantiates that the publication output of more productive authors exhibit also higher mean citation rates of their output. We have to emphasise that this holds at least for registered authors.

Table 2. Indicators of productivity and citation impact of RID authors and their papers by major science fields. [Data sourced from Thomson Reuters Web of Science Core Collection].

Field	y	z	x	f_0	g_0
A	2.76	1.32	3.65	17.9%	26.4%
Z	3.26	1.40	4.57	14.6%	22.8%
B	4.85	1.19	5.78	11.6%	15.9%
R	3.56	1.15	4.10	16.7%	22.0%
I	4.83	1.58	7.62	13.0%	20.9%
M	3.01	1.76	5.29	16.8%	27.8%
N	3.75	1.54	5.78	14.0%	20.8%
C	4.27	1.89	8.08	12.9%	20.8%
P	3.56	1.66	5.91	14.7%	25.1%
G	3.85	1.40	5.39	15.1%	20.9%
E	2.19	1.35	2.96	26.0%	36.4%
H	1.52	1.47	2.23	35.5%	44.6%

[*] Legend: A: agriculture & environment; B: biosciences (general, cellular & subcellular biology; genetics); C: chemistry; E: engineering; G: geosciences & space sciences; H: mathematics, I: clinical and experimental medicine I (general & internal medicine); M: clinical and experimental medicine II (non-internal medicine specialties); N: neuroscience & behavior; P: physics; R: biomedical research; Z: biology (organismic & supraorganismic level)

In order to conclude the analysis, we have calculated the mean values of the basic statistics x, y and z as well as the shares of cited authors and papers f_0 and g_0 by subject fields (see previous subsection for description). Table 2 shows these indicators for the 12 major fields in the sciences according to the Leuven–Budapest classification scheme (see Glänzel & Schubert, 2003). As explained in the theoretical part $x = y{\cdot}z$, $x \geq y$ and $f_0 \leq g_0$ is to be observed. Also subject-specific peculiarities are expected. The y and g_0 values concerning the citation impact of papers are by and large in line with the expectations: high impact and low share of uncited papers in the biomedical sciences and the opposite situation in engineering and mathematics. Nevertheless, the very high impact of chemistry (with low uncitedness) was somewhat surprising and somewhat deviates from the general citation patterns of the fields. Chemistry seems also to be somewhat overrepresented in terms of author registration; 33.5% of all RID authors are active in this field. This is followed by physics with 27.4% and biosciences with 20.8%. All other fields have shares of registered authors below 20% with neuroscience and mathematics having the lowest ones (7.6 % and 4.4%, respectively). In this context we have to mention that the distribution of shares of RID authors over fields is rather strongly correlated with the corresponding distributions of their papers ($r = 0.928$). Hence the question arises whether statistics as presented in Table 2 could be used as reference standards for publication activity and citation impact of authors at the national or institutional level. It has already be stressed in the introduction that an application at the individual level is not recommended because of the heterogeneous age and profile structure of the underlying reference data. Other details regarding this question will be tackled in the following subsection.

Limitations

After the methodological groundwork has been laid for capturing and describing the relationship between productivity and citation impact of authors and their papers, we have also to look at considerable limitations of possible applications of the indicators derived from this model. The low variation of average productivity over subject fields gives already a first hint of possible issues. As already observed by Heeffer et al. (2013) on the basis of the three-year publication period 2009–2011 and RID authors from eight selected countries, the distribution of average productivity was rather flat and ranged – except for physics – roughly between 2 and 3 papers by RID author. Only the average activity in physics with 5 papers per author was distinctly higher. The accustomed and specific inequality of citation impact of papers in different subject areas is almost missing in the productivity statistics what surprises since it is known that scientists in mathematics and engineering are usually less productive – at least as reflected by journal literature – than their colleagues in most fields of the natural and above all in the life sciences. The reason for the observed phenomenon is quite complex but readily explicable. In order to discuss this in detail we have first to refer to the corresponding statistics on citation rates of given paper sets. Provided that the publication year or period as well as the citation window is properly defined and chosen and the subject classification is appropriate, multiple subject assignment of individual papers is then the only severe issue to cope with. Various fractional counting and weighting models have been developed to overcome this problem and to build suitable reference standards for benchmark analysis. Even for more complex statistics than simple shares and means, fractionation by subject can still yield extremely robust statistics as the methods of characteristic scores and scales has shown for various citation windows and aggregation levels (cf. Glänzel, 2007; Glänzel et al., 2014). The question of co-authorship, in general, and how the individual co-authors' actual contribution to a paper should be credited, in particular, is at least in the context of paper-based citation indicators a secondary issue and not primarily related to the definition of citation indicators. The situation becomes completely different, whenever author productivity is directly included in indicator building as, for instance, in our "triangle model" based on the author-paper-citation relationship. The different (academic) age and the different profiles of authors have already been mentioned as possible sources of bias or even distortion, notably in the context of creating benchmarks for individual-author statistics. The most serious issues are related to co-authorship and cannot be simply solved by fractionation by co-authors and/or subjects. Collaboration of senior with junior co-authors, that is, of authors with strong publication record and less active authors, independently of their actual contribution to the paper in question and their function in preparing it, might have quite strong effect on the resulting indicators at the author level but also at higher level of aggregations. Here we would also like to point to two further issues, firstly the fact that a prolific author in one subject might only play a marginal part as researcher in a different subject in which he/she is collaborating with a possibly less prolific author, who, however, takes the part of the senior co-author of the paper(s) in this topic. Secondly, when it comes to measuring citation impact, an uncited author might be a co-author of a frequently cited author but the joint publications are not cited. This also implies that a mere author–citation analysis in conjunction with productivity studies does not yet suffice; an additional paper–citation analysis is needed for an adequate interpretation. And it becomes clear that a simple fractionation algorithm will not be able to solve these problems. A superposition of fractional counting at three levels (co-author credit, assignment by author profile and subject of publications) is required to solve at least the technical part of this problem: the large overlap by multiple assignments (authors, papers, subjects) could, of course, be resolved and indicators could then be additive over these actors and units at the price of very low robustness. Finally, the most important conceptual issue described in this subjection, the different roles of authors in different environments, will never be solved by using any algorithm.

Concluding discussion

Elementary statistics including relative frequencies and (conditional) mean values have been used to illustrate a simple model of the author-paper-citation relationship. Both opportunities and limitations have been sketched. The use of a joint model for studies of author productivity and impact at higher levels of aggregation is a topical issue in scientometrics: Hitherto the celebrated but also disputed h-index (Hirsch, 2005), originally proposed for the assessment of research performance at the micro level, was the only one that has combined these two aspects, and afterwards been extended for the use at higher aggregation level in the context of institutional and journal evaluation as well.

For illustration purposes, we have selected authors with ResearcherID and active in 2011 in order to exclude errors in author identification as far as possible. Of course, we have to mention that homonyms and synonyms still occur in RIDs too (cf. Heeffer et al., 2013) but the weight of errors is reasonably small. The main advantage of this model is the possibility of studying citation impact under the condition of the author's productivity, and the identification of high performance in terms of both productivity and impact. However, the same precision as experienced with "classical" citation indicators defined on paper sets could not be reached. The main problem is of conceptual nature: Authors and their papers might hold a different position in various environments created by co-authorship of subject-related issues. This has already induced Hirsch to revise his index in terms of co-authorship (Hirsch, 2010). His new indicator also substantiated that complex constellations cannot be described by separately fractionated parts of the model.

The conclusions drawn from this study are two-fold: On the one hand, author-identification systems need to extended in a reliable way to reach a nearly complete coverage of the author population in the database so that indicators based on author IDs can be considered representative enough to be used as reference standards. The limited discriminative power of author-based indicators and the heterogeneity of the underlying author population, on the other hand, prevents the use of the indicators for the analysis of individual research performance as well as in the context of fine-grained benchmark studies at higher levels of aggregations.

Finally, we would like to emphasise again the necessity and general use of the model introduced in this study, which is formally independent of any author-identification system. The model makes is possible to formalise and describe the relationship between authors, their publications and the citations those publications receive. The neglect of the structural properties and peculiarities of this “triangle relationship” might result in misinterpretation or even miscalculation of statistics and indicators at this level. The use of author identification in this context is an important means of demonstrating the measurement of this relationship for at least a considerable share of active authors.

References

Braun, T., Glänzel, W., & Schubert, A. (2001), Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics*, *51*(3), 499–510.

Caron, E. & van Eck, N.J. (2014), *Large scale author name disambiguation using rule-based scoring and clustering.* In: E. Noyons (Ed.), *“Context Counts: Pathways to Master Big and Little Data”. Proceedings of the STI Conference 2014, Leiden University, 2014*, 79–86.

Glänzel, W. & Schoepflin, U. (1994), A stochastic model for the ageing analyses of scientific literature. *Scientometrics*, *30*(1), 49–64.

Glänzel, W. & Schoepflin, U. (1995), A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, *21*(1), 37–53.

Glänzel, W. & Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Glänzel, W. (2007), Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, *1*(1), 92–102.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2009), Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, *78*(1), 165–188.

Glänzel, W., Thijs, B., & Debackere, K. (2014), The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*, *101*(2), 939–952.

Heeffer, S., Thijs, B., & Glänzel, W. (2013), Are registered authors more productive? *ISSI Newsletter*, *9*(2), 29–32.

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*(46), 16569–16572.

Hirsch, J.E. (2010), An index to quantify an individual’s scientific research output that takes into account the effect of multiple co-authorship. *Scientometrics*, *85*(3), 741–754.

Moed, H.F., van Leeuwen, T.N., & Reedijk, J. (1998), A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation*, *54*(4), 387–419.

Price, D.D. & Gürsey, S. (1976), Studies in scientometrics. Part 1. Transience and continuance in scientific authorship. *International Forum on Information and Documentation*. *1*, 17–24.

Schubert, A. & Glänzel, W. (1984), A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, *6*(3), 149–167.

Schubert, A. (2007), Successive h-indices. *Scientometrics*, *70* (1), 201–205.

Strotman, A. & Zhao, D. (2012), Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820–1833.

Tang, L. & Walsh, J.P. (2010), Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3), 763–784.

Thomson Reuters (2012), Web of Science® Help. Accessible at: http://images.webofknowledge.com/WOKRS58B4/help/WOS/hp_das1.html. Last modified on 09/18/2012, accessed on 28/12/2014.