Developing a model for individual-level measurement of research activity and citation impact

Glänzel, W.; Thijs, B.; Zhang, L.; Sivertsen, G.

STI 2018 Leiden

23rd International Conference on Science and Technology Indicators

"Science, Technology and Innovation Indicators in Transition"

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

Developing a model for individual-level measurement of research activity and citation impact

Wolfgang Glänzel[*], Bart Thijs[**], Lin Zhang[***] and Gunnar Sivertsen[****]

[*] *Wolfgang.Glanzel@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary

[**] *Bart.Thijs@kuleuven.be*
KU Leuven, ECOOM, Leuven (Belgium)

[***] *Zhanglin_1117@126.com*
Wuhan University, School of Information Management, Wuhan (China)
KU Leuven, ECOOM, Leuven (Belgium)

[****] *gunnar.sivertsen@nifu.no*
Nordic Institute for Studies in Innovation, Research and Education, Oslo (Norway)

Introduction

Advanced tools of citation analysis have laid the groundwork for various aspects of normalisation of citation-impact indicators at practically all levels of aggregation. Yet, normalisation and comparability remained an intractable problem, whenever publication activity formed a component of the indicators in question. The h-index, which depends on both number of papers published and citations received by those, might serve just as an example. In order to model such combinations, Glänzel et al., (2016) proposed a triangular model for publication and citation statistics of individual authors in order to study the possibility of creating reference standards for publication activity and citation impact of individual authors. Although the model met the theoretical challenges, its practical implementation definitely hit the wall for two particular reasons. The first reason we could identify was related to author identification. The use of researcher identifiers such as ORCID or the Clarivate Analytics Researcher-ID proved not representative enough to be used as universal standards. The same applied to the results of large-scale author-name disambiguation algorithms. The second issue resulted from a superposition of two effects: While citations can always be uniquely linked to the cited item, publication activity has, in the most cases, to be linked to several authors with different contributions to the papers and different profiles, who might also represent different subject matters. A bit simplifying, one could say that an author's productivity standard, which itself is subject to temporal changes, is not necessarily that of his/her co-authors. In the present study we will therefore make the attempt to analyse publication activity and citation impact on the basis of a unique cleaned national dataset, where the first factor, problems with large-scale author identification, is eliminated. The full individual validation of the set also helps gain clarity regarding the final limitations of the subject, profile and seniority factor in productivity studies. Research is done along the following questions.

- Do the features of clean national data allow the creation of a model for individual evaluation of research activity and citation impact?
- Can we extend the model of Characteristic Scores and Scales (CSS; e.g., Glänzel et al., 2014) to publication activity, if so what are the caveats?
- Does co-authorship fractionation matter, if so to what extent?
- How can we combine the publication CSS model with CSS classes for citation?

In order to answer these questions, a 10-year publication dataset of the Norwegian Cristin database (cf. Sivertsen, 2016) was used in combination with citations from the Clarivate Analytics Web of Science Core Collection (WoS).

Data sources and processing

All papers recorded in the Norwegian Cristin database from Norway's four largest universities (Bergen, Oslo, Trondheim, Tromsø) and indexed in the 2005–2014 WoS journals databases were selected. Citations were collected within three-year windows for all papers of document type article and review. All Cristin data are fully validated for correctness and were provided as anonymised dataset. Author names were not provided.

Methodological aspects

In order to answer our research questions, we have applied the following methods. For the "productivity" of Norwegian researchers we have calculated baseline profiles by discipline. For doing so, we have used the 74 sub-field classification according to the refined Leuven-Budapest scheme (Glänzel et al., 2016). We have applied both full and fractional counting (by number of co-authors) and calculated the activity distributions over authors by discipline and compared the results. In a second step, we have calculated the CSS citation classes by the authors' *papers*. While the first set of CSS classes could be determined on the national basis only, we could apply the complete WoS as the baseline for the last step (see Glänzel et al., 2014).

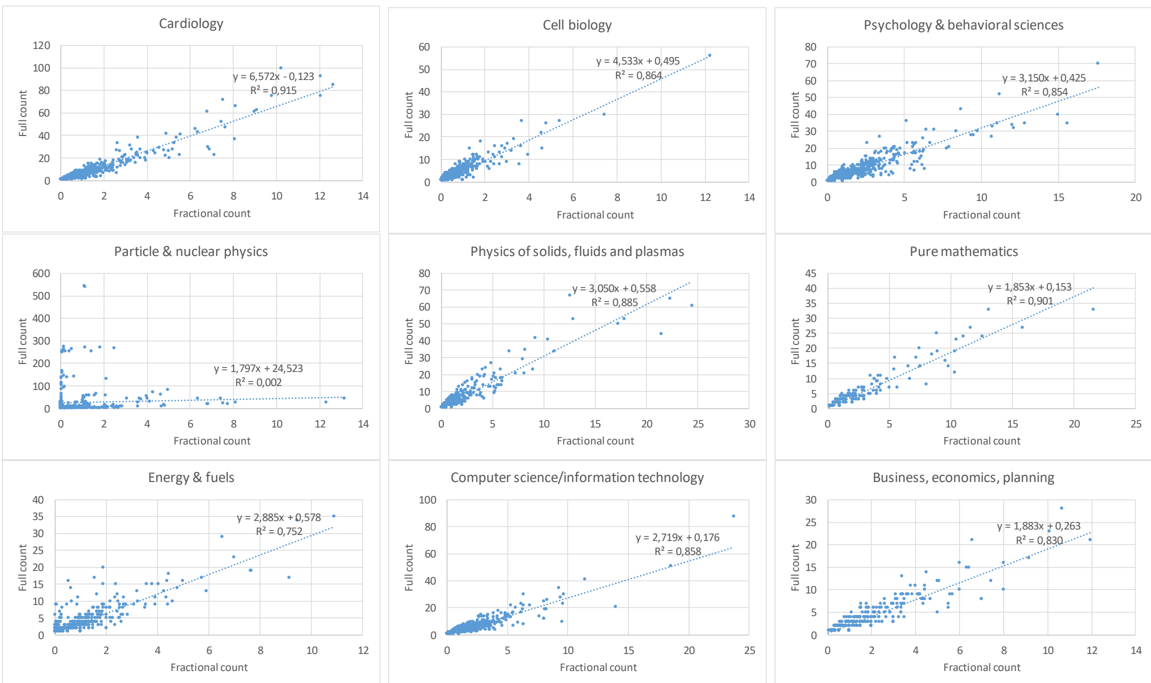
Results

Publication activity

In our previous, more methodological study (Glänzel et al, 2016), we sought to find reference distributions as baselines for individual authors' publication activity. We proceeded from the assumption that this could be solved analogously to citation distributions of publication sets or samples, once individual authors could be uniquely identified and assigned to subject profiles. This assumption turned out to be more problematic than expected. Firstly, we can define an author's publication profile using the cognitive subject assignment of the author's publications. Yet, just to use an analogue from citation analysis, this would be the same as if we had to base the citation standard for a publication on the basis of the citing papers. This approach increases the spectrum of subject assignment, notably in an interdisciplinary environment. The second issue is a result of co-authorship. The number of co-authors considerably differs across fields with rather low numbers in mathematics and social sciences and rather large numbers in the natural and life sciences, let alone for the sometimes- exceptional large number in astronomy, astrophysics and high-energy physics. In order to formally compensate the effect of co-authorship, we have fractionated publication counts and used this approach along with full publication counts. However, beyond the mere numbers, there is also a conceptual issue, particularly, the individual profiles of co-authors with their specific publication behaviour and the different "seniority" of researchers co-authoring the same paper. As a consequence, an author might appear to "take a low profile" in a subject A although the same researcher is very active in another one (say subject B). Also, the constitution of co-authors in A might be different from that of B. This implies that the two profiles (A and B) should not be merged. Just to give an example, one of the authors of the present study had to adjust to the practices and editorial policies of the life sciences when publishing with co-authors in their field, which partially resulted in papers that are otherwise not typical of the author's profile. Because of such field-specific publication behaviour we have to analyse productivity distributions by disciplines separately. In order to obtain reliable results, we have chosen the sub-field level, which allows to distinguish between the different standards, for instance, of mathematical & theoretical physics and particle & nuclear physics. On the other hand, the granularity is coarse enough to minimise the effect of outliers and to obtain sufficiently large sample sizes for the respective reference standards.

In a first step, we have looked at the effect of the number of co-authors on activity counts. Since individual publications were assigned to disciplines, authors can be assigned to more than one discipline. We have selected nine out of the 74 sub-fields according to the already mentioned Leuven-Budapest scheme. We have analysed more than these nine disciplines, but the selection representing several fields can stand pars pro toto for all fields in the sciences and social sciences. The selected fields are cardiovascular & respiratory medicine (I1), cell biology (B2), psychology & behavioral sciences (N2), particle & nuclear physics (P5), physics of solids, fluids and plasmas (P6), pure mathematics (H2), energy & fuels (E3), computer science/information technology (E1) and business, economics, planning (L1). A simple linear regression analysis shows a generally strong correlation between full and fractional-count based on publication activity. The regression plots are shown in Figure 1. We encountered two remarkable exceptions, particularly, particle & nuclear physics (left hand chart in the second row) and similarly, astronomy & astrophysics (not plotted here). It is worth mentioning that by far not all papers in particle & nuclear physics reflect "hyper-authorship" since we also found many papers in this discipline that have rather a low number of co-authors. This explains the observed polarisation in the chart and the un-correlatedness (R^2 = 0.002).

Figure 1: Plot of full count vs. fractional count of Norwegian authors' publication activity by selected disciplines.



In a second step we have calculated the scores for the CSS classes over publications. There is an essential difference between "productivity" and citation-impact CSS. Reliable data on publication activity apply to non-zero activity as information on authors who are temporarily or permanently not active is usually not available. Nevertheless, zero-truncated Paretian distributions can readily be integrated in to the underlying model (e.g., Telcs et al., 1985) and there is no effect on the resulting class distribution. The CSS scores and classes (four classes each, from CCS1 for low productivity to CSS4 for outstandingly active) are given in Table 1 for both full and fractional co-authorship counts. The observations can be summarised in three groups. The interpretation of the first observation is straightforward. The scores for the fractional counts are lower than the corresponding full counts and the deviation is in line with the communication behaviour in the respective communities. Thus, the scores in mathematics (H2) and economics & business (L1) are still quite similar or at least of the same order of magnitude, while high-energy physics (P5) clearly reflects the effect of "hyper-authorship". Nevertheless, except for P5, the class distribution according to full and fractional counts are very similar. Hence, we can draw two conclusions, namely, that in fields like P5 with a high share of "hyper-authorship" and more authors in CSS class 4 than class 3 (4.5% > 2.1%), individual publication activity analysis should not be conducted, neither using full nor fractional co-authorship counting. The second conclusion is rather positive: The CSS class distribution proved robust and independent of the actual counting scheme, of course, except for discipline P5. We therefore suggest the application of full counting. The last observation is the most critical one. We observed in all disciplines that the distribution deviates from the standard rule found in the context of citation impact: 70%–21%–6.5%–2.5%. The shares in the lower two classes considerably differ from the citation model. The explanation for the surplus in CSS1 can be found in the occasional activity of authors outside their main profile. Therefore, the model appears to be useful in the identification of more prolific authors (CSS3&4) in the disciplines under study. Furthermore, b_1 score gives the national mean activity in the respective discipline. Finally, we have to stress that the results presented here have to be interpreted in the national context with the Norwegian productivity baseline although international co-authorship may also have effect on national publication activity.

Table 1. CSS scores and classes for nine selected disciplines based full and fractional publication count.

Field	Full count			Fractional count			Full count				Fractional count			
	b_1	b_2	b_3	b_1	b_2	b_3	CSS1	CSS2	CSS3	CSS4	CSS1	CSS2	CSS3	CSS4
I1	3.91	9.74	22.79	0.61	1.64	3.33	74.9%	17.1%	5.4%	2.6%	76.9%	15.6%	4.8%	2.7%
B2	2.18	5.51	9.76	0.37	0.93	1.74	77.8%	15.4%	4.8%	2.0%	74.1%	18.2%	5.9%	1.9%
N2	3.25	9.43	17.25	0.90	2.47	4.69	77.3%	15.4%	4.8%	2.5%	74.1%	17.7%	5.7%	2.5%
P5	26.06	113.21	241.50	0.85	2.24	4.72	79.8%	13.6%	2.1%	4.5%	69.8%	21.9%	5.5%	2.9%
P6	3.51	10.75	23.03	0.97	3.00	6.78	78.1%	15.7%	4.6%	1.6%	77.6%	16.3%	4.5%	1.6%
H2	3.59	9.63	17.71	1.85	4.75	9.22	74.5%	16.7%	5.0%	3.8%	72.3%	19.5%	4.7%	3.5%
E3	2.46	6.19	12.07	0.65	1.61	3.14	75.3%	17.7%	4.6%	2.4%	72.1%	19.8%	5.5%	2.6%
E1	2.54	6.54	12.74	0.87	2.22	4.24	75.3%	17.4%	4.9%	2.4%	73.2%	18.5%	5.8%	2.5%
L1	2.32	5.89	9.42	1.09	2.67	4.55	76.5%	14.4%	6.6%	2.5%	73.5%	17.6%	6.1%	2.8%

Citation impact

The second part of this study brings us back to the well-known citation-based CSS model. Since this approach is based on papers, we could use the reference standard created by the complete WoS database. Scores are not calculated for each individual discipline separately but using fractionation and weighting by disciplines on the large-scale as described in Glänzel et al. (2014). For the complete Norwegian dataset we obtained the following distribution over CSS citation classes 1 through 4: 58.9%–28.1%–8.9%–4.0% This implies that the Norwegian WoS indexed publications form a sample that is biased towards higher-than-standard impact (with distinctly less poorly cited and more highly cited papers, cf. national statistics in Glänzel et al., 2014, 2018). In order to guarantee fair benchmarking, we compare individual citation impact with the national standard. In previous studies we have found five paradigmatic deviations of individual CSS-class distributions from reference standards (Thijs et al., 2017; Glänzel et al., 2018). These correspond to combinations of higher and lower shares of poorly and highly cited papers. No distinct deviation can be considered being in line with the baseline distribution (Type III). Columns 2–4 of Table 2 summarise the criteria. "0" means no distinct deviation from the standard, "+" ("–") means greater (less) than the expected share (Norwegian baseline). The most frequent and significant patterns form five main types. In the above-mentioned two papers, we have determined individual profiles for authors with at least 20 publications and applied a χ^2-test. However, the validity of this test proved to be somewhat problematic as it becomes very sensitive for deviations from the reference or expected profile with higher number of publications. It turned out that any prolific author had a profile that differs significantly and that classification of authors with the same profile was often just a matter of the number of publications. In order to obtain more robust results, we multiplied the baseline shares with 0.9 and 1.1, respectively, to obtain margins for considering the deviation. In addition, we selected only authors with at least 30 publications in the 10-year period, independently of their subject field. This consideration is based on statistical aspects, even if this results in underrepresentation of researchers in fields with typically lower publication activity as, for instance, mathematics or engineering. Nonetheless, 4.4% of all registered Norwegian authors met this criterion. In this context we have also to mention that not all Norwegian authors were active in the complete period and on the basis of the anonymised data we have no further information about "newcomers" and "terminators" (cf. Price and Gürsey, 1976). The selection resulted in an interesting sample that illustrates the distribution of all possible individual profiles. The two extreme profiles (Type I – "all around the moderately cited" and the polarised Type V – "everything but the moderately cited") were rather the exception to the rule. About 30% were in line with the national standard, while about one quarter had a less advantageous and 40% more advantageous profiles. The selection of more prolific authors thus resulted in a certain bias towards Type IV. Construction of and distribution over profile types are given in Table 2.

Table 2. Construction of profile types and distribution of most prolific Norwegian authors over profile types.

Type	**CSS1**	**CSS2**	**CSS3&4**	**Share of authors**
I	–	+	–	5.9%
II	+	0	–	23.4%
III	0	0	0	29.6%
IV	–	0	+	40.1%
V	+	–	+	1.0%

Conclusions and discussion

Our study demonstrates that statistics on the publication portfolios of individual authors needs to be studied in the context of their career. This information was not available for the present study and will be added in a follow-up study. Nevertheless, we were able to summarise individual-level productivity and citation indicators in a large dataset in a meaningful way. To answer the research questions, we obtained surprisingly interesting results. The correctness of author assignment proved the most important issue in creating individual research statistics at the national level. This also avoided the biases observed in previous studies, for instance, in the context of researcher IDs. The features of clean national data allowed the creation of a model for individual evaluation of research activity and citation impact, which can even be enriched as we add career specific information at a later stage.

Nevertheless, not all issues could be resolved since those are rather of conceptual than of technical nature. In the very first place, we have to mention co-authorship, which is not merely a numeric problem that could readily be solved by fractionation. On the contrary, except for subjects that are "notorious" for hyper-authorship, the influence of the number of co-authors on activity is rather limited and not unambiguous either (cf. Fig. 6 in Braun et al., 2001). This is also in line with previous observations, namely the constitution of co-authorship teams in terms of seniority and subject profiles. Hence it does matter if an author acts as junior/senior partner in different environments and communities. Also, international co-authors, who are not part of the national statistics, but nevertheless do influence national publication activity and citation impact, play an important part although they do not appear in the national statistics themselves. This also answers the second and third question addressed in the introduction. The CSS model can be extended to publication activity even if the distribution is truncated from the left at point zero, but the caveats lie in the nature of research collaboration, communication, individual stages of academic careers and subject-specific socialisation. The role of the actual number of co-authors is, however, secondary. The CSS model can furthermore be used to set standards for the identification of highly productive authors within given disciplines and to obtain clean performance classes for individual authors. Here, just as in all evaluative exercises on individual researchers, the sample size sets the only severe limitation. The selection of researchers with sufficiently larger number of publications, however, results in the well-known built-in bias towards higher citation impact. The present study could confirm this effect as well. In due consideration of this effect, the bibliometric method proposed in this study can indeed advance the measurement of individual-level research activity and citation impact. In a follow-up study we will apply the same methodology to Flemish research on the basis of similarly clean author data.

Acknowledgment

Lin Zhang acknowledges the support from the NSFC of China (Grant No.71573085). Gunnar Sivertsen acknowledges the support from the FORINNPOL programme, project number 256223, at the Research Council of Norway.
Data displayed in Tables and Figures are partially sourced from Clarivate Analytics Web of Science Core Collection.

References

Glänzel, W., Thijs, 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.

Glänzel, W., Heeffer, S., Thijs, B. (2016). A triangular model for publication and citation statistics of individual authors. *Scientometrics*, 107(2), 857–872.

Glänzel, W., Thijs, B., Chi, P.S. (2016). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: The Book Citation Index. *Scientometrics*, 109(3), 2165–2179.

Glänzel, W., Thijs, B., Debackere, K. (2018). *Citation classes – a distribution-based approach to profiling citation impact for evaluative purposes*. In: W. Glänzel, H.F. Moed, U. Schmoch, M. Thelwall, Springer Handbook of Science and Technology Indicators. Springer – Heidelberg, in press.

Price, D. de Solla, Gürsey, S. (1976). Studies in scientometrics. Part 1. Transience and continuance in scientific authorship. *International Forum on Information and Documentation*, 1(2), 17–24.

Sivertsen, G. (2016). Data integration in Scandinavia. *Scientometrics*, 106(2), 849–855.

Telcs, A., Glänzel, W., Schubert, A. (1985), Characterization and statistical test using truncated expectations for a class of skew distributions. *Mathematical Social Sciences*, 10(2), 169–178.

Thijs, B., Debackere, K., Glänzel, W. (2017), Improved author profiling through the use of citation classes. *Scientometrics*, 111(2), 829–839.