The P-model:
An Indicator that Accounts for Field Adjusted Production as well as Field Normalized Citation Impact.

Erik Sandström[1], Ulf Sandström[2] and Peter van den Besselaar[3]

[1]erik.sandstrom@zoho.com Stockholm (Sweden)
[2] ulf.sandstrom@indek.kth.se KTH Royal Inst Technol, Stockholm (Sweden)
[3]p.a.a.vanden.besselaar@vu.nl Vrije Universiteit, Amsterdam (the Netherlands)

Abstract

Any type of scientific study or evaluation of research quality and impact enters into two types of problems if there is more than one topic area involved in the study: (1) How to account for differences in (paper) production? (2) How to account for differences in citation impact, i.e. influence over subsequent literature? This paper aims to show that these questions can be answered with the help of two methods; the Field Adjusted Production (FAP) indicator and a percentile indicator which is designed to include the FAP. Consequently, they are used in combination in order to express a score that includes both paper production an impact into one figure. Thereby is constructed a score that can be used for ranking of universities, departments, individuals. The paper first explains the background of the method, and then how to calculate the indicators belonging to the P-Model. Then the paper indicates some examples and will discuss methods for validation of the proposed indicator.

Introduction - a long discussion

Performance indicators seem to be subject of continued interest for the bibliometric community. After a period of 10-15 years doing good work with quite some interest from the professional society of research administrators, the fall of the crown indicator has stimulated renewed interest and a critical discussion of bibliometric indicators. Suddenly it was (re)-discovered that mean values was not the best way to handle bibliometric data. Re-discovered because it was already indicated a long time ago by Paul R McAllister, Francis Narin and James G Corrigan in a paper with the title, "Programmatic Evaluation and Comparison Based on Standardized Citation Scores" (IEEE Transactions on Engineering Management 1983)[1].
Their indicator uses a transformation to the logarithm of the number of citations (plus one-half to include the zero-cited papers) and measures this in standard deviations from the mean. This means that many of the ingredients of the indicator discussion that has taken place since Lundberg (2007) already have been available a good time before the Leiden and Leuven indicators were created and launched in Europe (Moed & van Raan 1988; Schubert, Glänzel & Braun 1988). This cultural divide seems even more idiosyncratic as the National Science Indicators in the US have used percentiles since long time ago, but in Europe, they didn't gain interest as the so-called *crown indicator* had such a strong market position. It took a situation where the trust or confidence in indicators had reached its bottom (Wilsdon et al. 2015, Hicks et al 2015) before the percentiles came into the European discussion. After that first wave of the European discussion pioneered by Leydesdorff and Bornmann (in several papers) it is now time to start building stronger and more sustainable indicators.

This paper presents a composite indicator called the P-model which combines production *and* impact into one score and is size-dependent to its nature. The act of combining papers and citations has been done before e.g by the Leiden group (P time MNSC) but with important problems on both sides of the multiplication. Here we suggest solutions to these problems.
We will evaluate the P-model using criteria suggested by Yves Gingras in the edited book *Beyond Bibliometrics* (ed. Sugimoto 2014): 1) Adequacy; 2) Sensitivity and 3) Homogeneity.

[1] The paper is accessed at <https://www.forskningspolitik.se/files/dokument/programmatic-evaluation-and-comparison-based-on-standardized-citation-scores.pdf>

Field Adjusted Production – Waring distributions

Field differences in production are well known; medical researchers tend to produce more, often shorter papers where methodology and prior knowledge is codified in citations. Engineering scientists are known to produce less frequently and have fewer cross-references (Narin and Hamilton, 1996; Glänzel, 1996). These field differences affect both citation rates and the number of papers per author, differences that are to some extent explained by the shifting coverage of publication activity in the WoS database.

Let us say that we want to stay with the WoS database due to its good features: selection of sources, prudence, etc. How do solve the problems? In order to compute a field adjusted factor, we have to get rid of certain obstacles: publication databases give information on the authors that are active during a given period, not all the potential authors. As the non-contributors (non-publishing authors) are unknown it is difficult to calculate an average publication rate per author taking all potential authors into account. But, there is a proposed mathematical solution to this problem: bibliometric data are characteristical "Waring distributions" (Schubert and Glänzel, 1984). Using information on the distribution of author publication frequencies an estimate of the average publication rate per researchers (contributors and non-contributors) in a given field and country can be computed (Telcs, Glänzel & Schubert, 1985).

The approach is based on mathematical statistics and a theoretical discussion can be found in papers by Braun, Glänzel, Schubert & Telcs during the second half of the 1980s. Inspired by Irwin (1963) they showed that bibliometric material had the properties of "Waring distributions". A straight line should be obtained by plotting the truncated sample mean of these distributions (Telcs, Glänzel & Schubert, 1985). By extrapolating this series to Origo, the numbers of non-contributors are included. The intercept of this line is the average productivity of all potential authors during a given period of time (Braun, Glänzel & Schubert, 1990). In our model, this value is used as a reference value and is computed per field for Nordic data. Several successful empirical tests using the Field Adjusted Production (FAP) model have been implemented (e.g. Schubert and Glänzel 1984; Schubert and Telcs, 1986; Buxenbaum, Pivinski & Ruberg, 1987; Schubert and Telcs, 1989; Sandström and Sandström, 2008b). A more complete article on this method was published by Koski, Sandström & Sandström (2016). Here we follow that latter source for the explication of the method.

The Field Adjusted Production is calculated as follows:

$$\sum_{i=1}^{n} \frac{\boldsymbol{P_i}}{\boldsymbol{r_i}}$$

where P_i is the number of papers in field i and r_i is the (estimated) average number of papers per researcher in field $_i$. The estimation of the reference values is performed for each field by first calculating the s-truncated sample mean of each field as follows:

$$\frac{\sum_{i=s}^{\infty} \boldsymbol{in_i}}{\sum_{i=s}^{\infty} \boldsymbol{n_i}}$$

Where n_i is the number of authors having exactly i papers. The truncated sample means are plotted versus s and the intercept of the fitted line, using weighted least squares linear regression, is used as an estimate the number of papers per author for the entire population The regression is weighted using proposed method for that by Telcs et al. (1985).

When applying this model, authors with an address at Nordic universities are used as data. Homonyms and similar problems are taken care of by automatic procedures in combination with manual procedures. This was done for all Nordic universities (Sweden, Finland, Denmark, and Norway) and the operation yielded almost 400,000 unique authors for the period 2008–2011.

Field delineation is maybe the most important issue here. The Thomson/ISI subject categories are used for citations, but these some 260 categories create too small samples when Nordic authors are used to constructing the productivity data. There are several alternative ways of producing macro classes (e.g. the Clarivate ESI 22 field categories). We have been using journal inter-citations as proximity values (Boyack and Klavans, 2006), and with the least frequent relation as decisive in order to distinguish, as far as possible, between basic and applied sciences. It has been shown by Rinia, van Leeuwen, Bruins, van Vuren and van Raan (2002) that applied sciences tend to cite back to more basic sciences, not the other way around. But the clustering procedures that were tried didn't really work as good as we wanted and therefore we decided, after some reiterations, that the suggested macro fields in the Science Metrix classification would fulfill the requirements we knew where needed, e.g. to distinguish a category of applied science fields. So, we had in the final round five different clusters (fields); humanities, social science and economics, applied sciences, health sciences, and natural sciences.

The methodology described was used to establish a reference value based on disambiguated researchers from all Nordic universities. By using the count of paper fractions per author and relate that to the reference value (the field factor) we obtain the relative quantity of production performed by the person or the unit (the indicator is further explained below, see Table 2). This indicator is called the “Field Adjusted Production (FAP)” and can be explained as the expected production per area over a period of time (in this case a four year period) and for a “normal” researcher with all other assignments at the same time. One can say that the indicator expresses how many persons the actual production score accounts for, if the value is ten for a group of people, then that can be related to the actual number of people in the group. So, if they are five and they publish in the range of ten persons then the production is 100 % higher than what would be expected.

Citation impact – percentile distributions

The literature on citation impact is wider and more diverse than the one on productivity but much of it is somewhat dated and irrelevant (Waltman 2016; Abramo 2018). There are three major questions that we will touch upon before we present the methods that were applied in this project. 1) What do we mean with the term “citation impact”; 2) Percentiles instead of averages; and 3) Size-dependent vs. size independent indicators.

The discussion on citation impact from research has intensified and has been widened over that last ten years (Bastow et al. 2014). Opening the concept of impact to all types of influence on society has many advantages in the dialogue with politics and funding agencies but at the same time the concept a bit vague. Therefore, it should be possible to talk about two different concepts of impact, the first one is the restricted impact and the second one is the wider and looser concept of impact. Depending on the fact that this exercise is a quantitative study of the relation between gender diversity and research performance we lean towards the first version of the concept of impact which is neatly laid out by Abramo (2018).

In the understanding of Abramo (2018) a paper might have an impact on the subsequent literature and for this, he reserves the concept “citation impact”. It follows that we can use a very precise measure based on how articles are cited even if it should be considered as a proxy as the reference behavior is an non-harmonized process, many different types of behavior are detectable, but in the long run and with large stocks of papers there should be possible to use statistical methods that do not suffer from the noise in the data.

Calculation methods built on averages are of less interest as we can easily understand that there are drawbacks with methods that measure impact as a mean of all papers over a period of time. When the same author publishes another paper, the first papers' impact does not disappear or diminish. On the contrary, it can be made stronger by new evidence. Therefore, the overall impact of the two papers cannot be measured by an average, and instead, an additive method is required. In order to proceed with that method, we apply the FAP score introduced in the former section and illustrated in detail below.

Instead of averages, the method for performance analysis is partly based on a percentile approach. All articles in each group of articles are ranked based on citations. The field is defined according to the subject categories specified in the Web of Science database, and the articles are divided into percentile classes, the top 1% (99th percentile), 5 %, 10 %, 25 %, 50 % and below 50 %. Measures based on percentiles have the advantage of not being affected by causes of bias in citation distributions (Rousseau 2005). In certain disciplinary areas, a few publications with very numerous citations otherwise boost the mean, which can result in 70% of the articles in the area being below this mean (c.f. Campbell 2017 [STI 2017 paper], c.f. Thelwall 2019).

With this, we turn to the Percentile Model and how it allocates points for each article. The points are based on probability. An article that is among the most highly cited 1% of articles is assigned 100 points; one in the top 5 % is given 20 points and so forth (see Table 1). An article that is among the 50 % least cited is given 1 point, which means that a researcher can never lose from getting an article published. The points thus received by each article are then corrected by the field-adjusted production (FAP) method to compensate for differences between research areas in the rate of scholarly production. Such an approach provides a lot of information and should be useful to summarise performance in a single value. The method is preliminary called P-model or the Influence Factor.

Table 1. Points allocated per percentile group in the P-model

Percentile group	Points
TOP1 % (99th percentile)	100
TOP5 %	20
TOP10 %	10
TOP25 %	4
TOP50 %	2
TOP100 %	1

Note: Based on Sandström & Wold (2015)

The idea to allocate points is inspired by Leydesdorff (2012) and Leydesdorff & Bornmann, (2011). They suggested the following: *"(...a method to) calculate a mean of the ranks weighted by the proportion of papers in each. The minimum is 1, if all papers are in the lowest rank; the maximum is 6 if they are all in the top percentile."* Although of interest this method is dubious as the groups are not of the same size.

One major problem with the points in the P-model (see Table 1) is that there is quite a large difference between top1% and top2% which has been pointed out as reactions to evaluations based on this model (Henreksson, personal communication). Pragmatic reasons are to a large extent behind the model as it is dependent on the programming of the indicators and the calculations would be too extensive for a normal personal computer of today.

The model ingredients in detail

Now we have the two components of the model and in the following, we will go through it in detail so that if the reader would have the two different basic calculations done (FAP and P-model) he/she should be able to finalize it to a score. Here is the method for calculation and the text refers to Table 2 below:

(1) Each publication is from a source (SO), the periodical (journal) name.
(2) REF stands for reference value based on Nordic values; the first line is about 0.86 which means that an article by one author alone would account for more than what would be expected from one author in the Nordic countries, 1.0 is the expected value.
(3) Then Frac P showing the fraction of a paper or how many authors were involved in the production, in this case, four authors.
(4) This is transformed to a FAP value of 0,29 (i.e. Frac P/REF). This indicates that over a four-year period an author is expected to publish about four such papers.
(5) Then follows six different columns giving information on the percentile fraction, a fractionalization based on the fact that we are working with integers and therefore there can be many with a same number of citations at the border of a percentile group. An elegant solution to this is fractional counting suggested by Waltman and Schreiber (2013).
(6) Based on that there are six FTOP columns which give the product of the calculation [FAP*(Fraction) Percentile Group]*P-Model points(100; 20; 10; 4; 2; 1).
(7) Total sum in the column to the right gives the total per article. When totals are calculated they can be summarised per person or per research team or any unit of interest.

Table 2. Example showing the calculation of the Percentile Model (P-Model) indicator

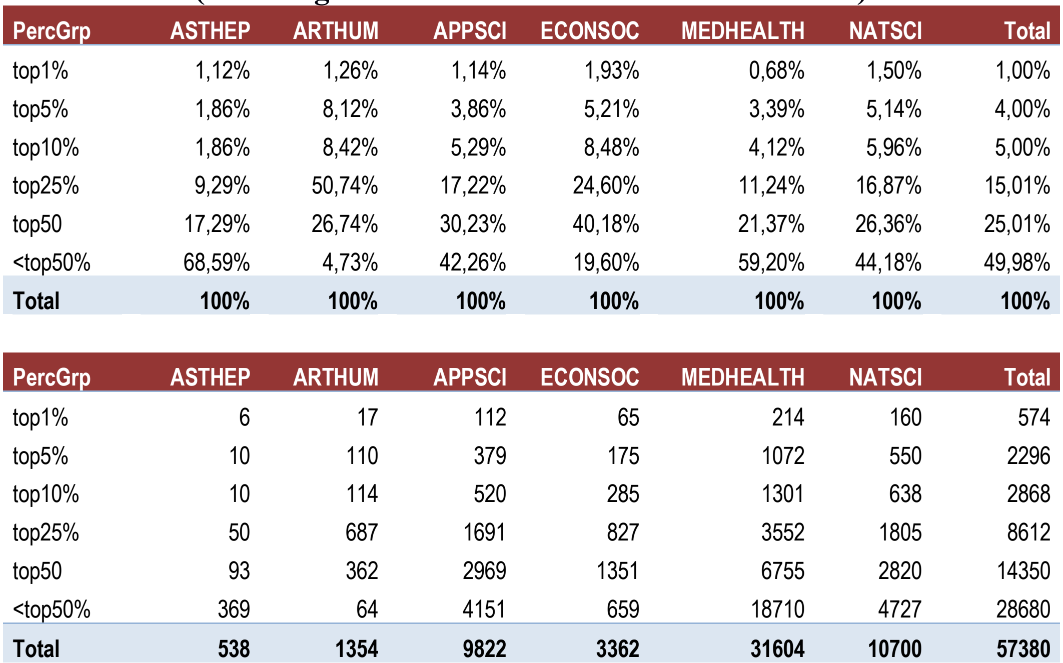
SO	REF	Frac P	FAP	TOP1	TOP5	TOP10	TOP25	TOP50	TOP100	FTOP1	FTOP5	FTOP10	FTOP25	FTOP50	FTOP100	P-Model points
LIGHTING RESEARCH & TECHNOLOGY	0,862958833	0,25	0,289700957	0	0,8	0	0	0	0	0	4,6352	0	0	0	0	4,635215314
JOURNAL OF CRYSTAL GROWTH	1,255448818	0,25	0,199131973	0	0	0,6667	0	0	0	0	0	1,32755	0	0	0	1,327547149
JOURNAL OF BIOSOCIAL SCIENCE	0,697673	0,333333333	0,477778749	0	0	0	1	0	0	0	0	0	1,91111	0	0	1,911114998
JOURNAL OF PROSTHODONTICS	1,171782937	0,333333333	0,284466792	0	0	0	0	0	1	0	0	0	0	0	0,284467	0,284466792

It should be remembered that we adhere to the principles of normalization to the field as has been practiced by the Leuven group (Glänzel et al. 1988) and implemented by the Leiden group (Moed & van Raan 1988). This is a central feature of bibliometric work: all types of performance, e.g. citation performance, are relative to the field where the object of evaluation has its publications. By publishing in a specific type of journals, authors tell the community of scientist that they want to be evaluated and measured by the standards in each subcategory of the fields available. There is some criticism towards the subject categories developed by Web of Science (see Leydesdorff 2008) or Scopus for that matter. But, the critique often fails to understand that a subject category are far from one-dimensional, instead, they include multi-assignations of each journal and it is, therefore, correct to say that there are thousands of categories in the WoS due to the multi-assignation methodology.

All calculations used here are based on three databases (SCI-E, SSCI, A&HCI) and four document categories only: Articles, Letters, Proceeding Papers, and Reviews. No other document categories are involved in the calculation of citation scores or the calculation of percentile groups. Author-based self-citations are deleted when the citation scores are calculated (based on the first author name).

Examples from the Swedish database showing how different areas are represented at every level of performance and that the indicator fulfils the criteria of equality between areas. But, there are obvious problems due to the differences between areas, e.g. Medical science are heavy in the bottom and social science is top heavy due to many low fraction authors in the former and full fraction authors in the latter domain.

Table 3. Fields distribution over Percentile Groups (disambiguated Swedish researchers - 2012-2015)



PercGrp	ASTHEP	ARTHUM	APPSCI	ECONSOC	MEDHEALTH	NATSCI	Total
top1%	1,12%	1,26%	1,14%	1,93%	0,68%	1,50%	1,00%
top5%	1,86%	8,12%	3,86%	5,21%	3,39%	5,14%	4,00%
top10%	1,86%	8,42%	5,29%	8,48%	4,12%	5,96%	5,00%
top25%	9,29%	50,74%	17,22%	24,60%	11,24%	16,87%	15,01%
top50	17,29%	26,74%	30,23%	40,18%	21,37%	26,36%	25,01%
<top50%	68,59%	4,73%	42,26%	19,60%	59,20%	44,18%	49,98%
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

PercGrp	ASTHEP	ARTHUM	APPSCI	ECONSOC	MEDHEALTH	NATSCI	Total
top1%	6	17	112	65	214	160	574
top5%	10	110	379	175	1072	550	2296
top10%	10	114	520	285	1301	638	2868
top25%	50	687	1691	827	3552	1805	8612
top50	93	362	2969	1351	6755	2820	14350
<top50%	369	64	4151	659	18710	4727	28680
Total	**538**	**1354**	**9822**	**3362**	**31604**	**10700**	**57380**

Note: ASTHEP is Astronomy & High Energy Physics; ARTHUM is Humanities; APPSCI is Applied Sciences; ECONSOC is Social Sciences; MEDHEALTH is Medical Sciences, and NATSCI is Natural Sciences. Upper table shows relative frequencies and lower table show raw numbers.

Further work

An important issue is how to validate the approach. Relevant validation criteria are:

1) Adequacy
2) Sensitivity
3) Homogeneity

These will be discussed in the presentation.