Vetenskapsrådet

Reference no.
113-2010-6148
Date
2017-01-16
Version
2.1

The bibliometric database at the Swedish Research Council - contents, methods and indicators

Abstract

Bibliometrics is used at the Swedish Research Council to study production volume, publication patterns and scientific impact. The underlying data for the database is bougth from Clarivate Analytics (previously Thomson Reuters) and correspond roughly to the content of *Web of Science*. Earlier the data have been delivered in a tagged data format but beginning in 2014, Clarivate Analytics are moving away from the tagged data format to delivering data as XML. In 2015 the Research Council switched to the new format, and at the same time changed the way some of the indicators are calculated. This document describes how the Research Council uses the data bought from Clarivate Analytics. It replaces the earlier document with the samte title, written by Kronman, Gunnarsson and Karlsson in 2010.

1 Introduction

The Research Policy department at the Swedish Research Council (SRC) maintains and develops a database for bibliometric analyses. This document describes the properties of the SRC bibliometric database, how the data is prepared, which indicators are used and how these indicators are calculated.

Section 3 describes the content of the database, Section 4 describes the calculation of the reference values that are used for the citation based indicators that are detailed in Section 5.

2 Indicator prerequisites

In most analyses at the SRC, each publication is divided into fractions based on the number of addresses on the publication and/or the number of subjects attributed to the issue in which the publication appears.

There are two different way of dividing a publication based on the addresses: *author-* and *organisation* fractionalization. Imagine a publication with 10 authors where 9 of them are from organisation A and 1 author are from organisation B. Using author fractionalization organisation A will be accredited 9/10 and organisation B will be accredited 1/10 of the publication. Furthermore, an author can have more than one address. Imagine that the author from organisation B is also affiliated with organisation C. Then organisation A still gets 9/10 of the publication but organisation B and C will get 1/20 each. With organisation fractionalization the three organisations in the example would be accredited with 1/3 of the publication each. For publications published before 2008 there does not exist a connection between the authors and addresses so these publications are divided using organisation fractionalization. For all publications that have a connection between the authors and addresses author fractionalization is used.

If a publication have 2 subjects and 3 authors where each author have 1 address it will be divided into 6 equally sized fractions, each having the weight 1/6 (1 divided with the product of the number of addresses and the number of subjects). If we want to count the publication volume based on subjects, each subject would be credited with one fraction, that is, 1/2 of the publication. If instead we want to look at the number of publications for different organisations and 2 of the 3 addresses comes from the same organisation, that organisation would be on the address of 4 of the 6 fractions and would hence be accredited with 2/3 of the publication. Furthermore, if we wanted to look at the publication volume for organisations based on subjects, this organisation would be accredited with 1/3 in each of the 2 subjects. These notions of fractions and weight will be used throughout Section 4 and 5. For a less technical description of the concepts described in Section 4 and 5 see [2].

3 Data source and properties

The SRC has an international publication database based on raw data bought from Clarivate Analytics. The content of the database approximately correspond to the content of *Web of Science* (WoS). This section describes the general properties of the raw data as it is delivered from Clarivate Analytics and the steps taken by the SRC to create a database and prepare the data for analysis.

3.1 The products from Clarivate Analytics

The products available from Clarivate Analytics can be divided into two collections: *Web of Science Core Collection (WOS)* and *Current Contents Connect (CCC)*. Each of these collections contains a number of different databases. All source records in WOS and CCC has a unique identifier called `UID`. The `UID` is a 15 character long string which is prefaced by the abbreviation of the collection from which the record is retrived. Example:

```
WOS:000345380400003
CCC:000282939200001
```

In the online version of *Web of Science* the `UID` is called the *Accession Number* (but more commonly known as the *UT Number*). The database at the SRC consist of databases from the WOS collection. Table 1 shows the different databases that the SRC database is based on, and the share of publications that are associated with each of them:

Edition	Name	Share of publications (%)
SCI	Science Citation Index Expanded	68.6
ISTP	Conference Proceedings Citation Index - Science	12.8
SSCI	Social Sciences Citation Inded	9.3
AHCI	Arts & Humanities Citation Index	7.7
ISSHP	Conference Proceedings Citation Index - Social Sciences & Humanities	1.0
IC	Chemical Indexes (Index Chemicus)	0.4
CCR	Chemical Indexes (Current Chemical Reactions)	0.2
BSCI	Book Citation Index - Science	0.0
BHCI	Book Citation Index - Social Sciences & Humanities	0.0

Table 1: The different products from Clarivate Analytics in the SRC database.

Each publication can belong to more than one product. The share of publications in Table 1 is based on fractionalised publications. If a publication belongs to both SCI and ISTP, 0.5 publications will be accredited to SCI and 0.5 publications will be accredited to ISTP. For analysis purposes each publication is reclassified as belonging to one of three categories: *Standard*, *Proceeding* or *Book*. The reason for this categorisation is that we want to make comparisons between publications of the same type, and the Standard and Proceedings categories often differ in the amount of data associated with each publication. Publications from the Proceeding category often only list the address of the reprint author which prevents address based fractionalisation (as shown in Table 2). Another issue with the proceedings is that they have a low citation count in general and that there are a lot of uncited publications which makes the field norms very unstable. See [2] for more on these issues. In general, only publications classified as Standard are used in the analysis carried out at the SRC. Even though the Proceedings publications are not counted in a normal analysis, the citations from the Proceedings publications to the Standard publications are counted.

The reclassification is done according to the following steps:

1. All publications that belongs to any of the products: SCI, CCR, AHCI, SSCI or IC are classified as Standard.

2. All publications not classified in step 1 which belongs to any of the products: ISTP, ISSHP are classified as Proceedings.

3. All publications not classified by the previous two steps are classified as Books.

Roughly 8% of all publications from the products SCI, CCR, AHCI, SSCI or IC are also classified as ISTP or ISSHP. The procedure above classifies these publications as Standard. The rationale behind this is shown in Table 2 which contains the share of publications in each product that has a certain number of author addresses. The distinguishing feature for publications that are only classified as Proceedings is the share of publications where the number of authors are 0 or 1. The distribution of the number of addresses for publications that belongs to products in both step 1 and step 2 (the Both-column) looks more like the distribution for the Standard publications and therefore they are classified as Standard.

Number of addresses	Standard Only (%)	Proceedings Only (%)	Both (%)
0	14	15	13
1	47	85	43
2	22	0	24
3	10	0	11
4	4	0	5
5	2	0	2
6	1	0	1
7	0	0	1
>7	1	0	1

Table 2: The share of publications in the different classes that has a certain number of addresses. (Data from Science Citation Index - Clarivate Analytics)

The database contains approximately 45 million publications from 1982 and onwards (the proceeding indexes starts 1989 and contains 6 million publications). Figure 1 shows the total number of publications in the database (starting from 1989 the proceedings publications constitutes about 16% of the publications). The database is updated in April each year after which it contains publications from 1982 until the previous year plus publications from the first quarter of the present year.

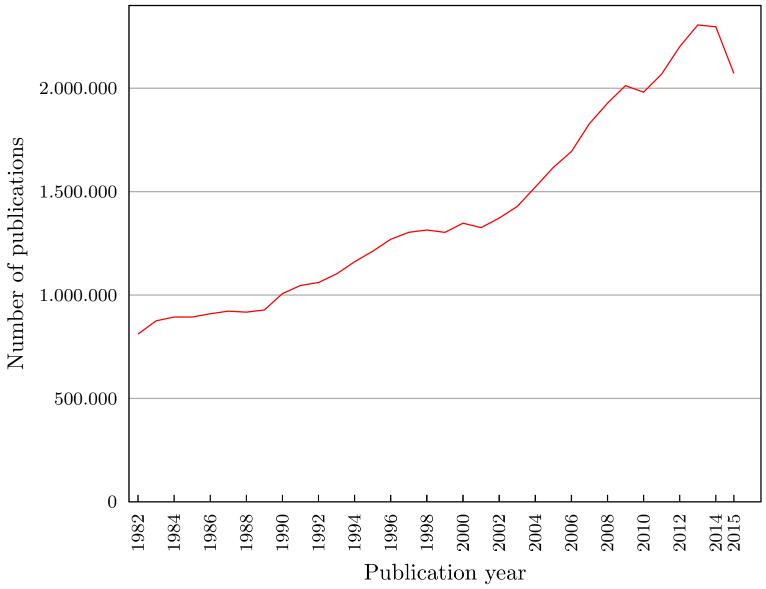

Figure 1: Number of publications in the SRC database. Starting from 1989 about 16% of the publications are Proceedings. (Data from Science Citation Index - Clarivate Analytics)

3.2 Document types

Clarivate Analytics classifies each publications as belonging to one of 39 different document types. Appendix A lists the document types and their corresponding share of publications as they appear in the data SRC receive from Clarivate Analytics.

About 4% of all publications have two documenttypes. Since the field normalisation is depending on the document type these publications are given one unique document type before the field norms are calculated. This is done in the following steps:

1. All publications classified as Standard where one of the document types are *Article* are classified as *Article*. (This step classifies 99% of the publications with two document types.)

2. For the remaining publications it appears that one of the document types are either *Book chapter* or *Book*. Since we do not have any other publications of these types these publications get the other document type.

Bibliometric analyses that includes citation level normalisations (i.e. comparisons to other publications of equal type) take the document type into account, since different document types tend to have different citation characteristics. But there are some issues in the way Clarivate Analytics identifies publications that are assigned to the document type *Review*. Two of the criteria for a publication to be classified as a *Review* is that there is a subsection called "Reviews" in the table of content, or that the phrase "review of the literature" are mentioned in the title, abstract or in the introduction.

The document type *Article* is a residual, that is, all publications that cannot be attributed to another document type are classified as *Article*.

One can easily find cases where the classification of reviews are wrong. One example are publications from the publisher *Annual Reviews* who publishes a number of series that only contains reviews of current topics. A lot of these publications are classified as *Article* by Clarivate Analytics. A rough estimate is that 10% of all *Review* publications are wrongfully classified as *Article*.

Most analyses performed at the SRC are based on publications with the document types *Article* and *Review*, but because of the problems mentioned with the classification of these document types the SRC does not differentiate between them when calculating the field norms.[1]

3.3 Retractions

In some cases Clarivate Analytics receives notifications from publishers about articles beeing retracted. In the database this is marked by adding "Retracted article..." to the title of the retracted article, and by adding "Retraction of..." to the title of the article that announces the retraction. At the Swedish Research Council, these articles are being removed from the database.

3.4 Journals

In the old data delivery format the data was organised around journal issues which meant that information about the issue, for example title, volume, issue number, where the same for all publications in the same issue of a journal. Each issue of a journal could also be identified by the first ten characters after the edition prefix of the UID. That is for the publication with `UID = WOS:A1992JA10900011` the issue

[1] In the old database SRC distinguished between *Articles* and *Reviews* and field norms were calculated separately for the two. Furthermore the *Article* document type previously also included documents of the type *Letter*.

number is `A1992JA109`.

In the new data delivery format each publication comes as one individual object and each publication element have child elements containing issue information.

To create journal statistics based on the new data format we group all issues that has the same `abbrev_11` title.

3.5 Subject areas

Each issue of a journal gets classified by Clarivate Analytics as belonging to between 0 and 10 subject areas. The publications inherits the subject classification from the issue in which they are published. In total there are around 250 subject areas.

Clarivate Analytics uses the subject area *Multidisciplinary Sciences* for journals, such as *PNAS*, *Science* and *Nature*, that contain papers from many different scientific areas. Since the field normalisation relies on the subject classification this means that publications in these journals are not compared to other papers of their "true" subject area, but rather to other publications classified as *Multidisciplinary Sciences.* This problem is aggravated by the fact that the *Multidisciplinary Sciences* journals often are prominent journals with high average citation levels. A publication in one of these journals is thus likely to receive a lower mean normalised citation rate than would have been the case if it had been published in a journal with a more specific subject attribution.

In an effort to find more article-relevant scientific subject classifications for publications in journals classified as *Multidisciplinary Sciences*, the publications in these journals are being re-classified by an algorithm using the publication's references and the (inbound) citations to the publication itself. The algorithm is based on the assumption that the subject areas of the publications referred to in reference lists and the citing publications indicate witch subject areas the referring/referred publication belongs to. The details of the reclassification is described in [1]. Using this reclassification method 60% of the publications in journals classified as *Multidisciplinary Sciences* can be classified as belonging to other subject areas.

3.6 Addresses

The addresses for the publications in the database are of two types: *author addresses* and *reprint addresses.* Normally a reprint address is not a author address but a special address that are used for correspondence with the authors of the publication. Before 1998 the address of the corresponding author was only registered in the database as a reprint address, but beginning in 1998 the address of the corresponding author is registered both as an author address and as a reprint address. This means that if we are interested in counting the number of organisations contributing to a publication we have to ignore the address of the corresponding author for publications published after 1997. But when back issues of journals are added or corrections are made to existing journals, addresses in publications published before 1998 are treated in the new way. To handle this, duplicate addresses are marked and excluded from the counting of contributing organisations. For reprint addresses the criteria for marking an address as a duplicate are:

- If the organisation is missing for the reprint address but not for the author address.
- Or if the `full_address` element of the reprint address and an author address are equal.

The criteria for marking an author address as a duplicate are:

- The `full_address` element are equal for two or more author addresses.

The counting of unique addresses are divided into three steps:

1. First publications published before 1998 are treated as above.
2. For publications published 1998 or later where the reprint address is the only address, this address is chosen.
3. For publications published 1998 or later that were not treated in the previous step, the author addresses are chosen.

3.7 Address mapping

After scanning the original address of a publication Clarivate Analytics alters the address in order for all addresses to have a similar structure. Some words are abbreviated[2] and the order of the words in the address is changed. For example, the following original address

Department of Cardiology, Karolinska Institute at Södersjukhuset, Karolinska Institute at
Karolinska University Hospital, Stockholm, Sweden

looks like this in WoS (and in the `full_address` element in the raw XML data)

Karolinska Inst Sodersjukhuset, Dept Cardiol, SE-11883 Stockholm, Sweden

Apart from the altered full addresses on a publication, the address data contains several fields where Clarivate Analytics tries to extract components of the address such as the organisation, postal code, street address, city and country. Depending on how the address is written on the publication the algorithms extracting these components can have trouble finding the right information. If we look at the address above we see that it contains two organisations: Karolinska Institutet and a Swedish hospital called Södersjukhuset. But when trying to extract an organization Clarivate Analytics combines the two into: Karolinska Inst Sodersjukhuset.

The SRC tries to identify addresses like the one above and divide them between the organisations contained in the original address. For addresses that does not seem to contain more than one organisation the organisation and city extracted by Clarivate Analytics is used to relate the address to a unified organisation name. Table 3 shows two examples of combinations of organisation and city that gets related to University of Gothenburg.

Organisation	**City**
Univ Gothenburg	Gothenburg
Univ Goteborg	Goteborg

Table 3: Two examples of combinations of organisation and city that gets related to University of Gothenburg.

These rules for attributing unified organisation names to addresses handles more than 99% of the Swedish publications in the database. For addresses from Norway, Finland, Denmark and Iceland a few Nordic collaboration projects have produced mappings between addresses and organisation names for the universities, university colleges and university hospitals in these countries.

There are similar rules for mapping different variants of country names to one of 249 standardised country names[3]. For instance, *Scotland* and *England* gets mapped

[2] Clarivate Analytics use around 340 abbreviations for common words like *Institute* which becomes *Inst* or *County* which becomes *Cty*.

[3] ISO 3166-1, https://www.iso.org/obp/ui/#search.

to *United Kingdom*, *Cambodia* and *Khmer Republic* gets mapped to *Cambodia* and *Fed Rep Ger* and *Ger Dem Rep* gets mapped to *Germany*.

In the source files delivered by Clarivate Analytics each publication record contains a list of author names and a list of author addresses. For records entered in the system before 2008 there is no indication of which addresses that relates to which authors, except for the corresponding author. As of 2008 most of the records delivered contains an indication of which author names and addresses belong together. A author and an address are linked via the `addr_no` field. All addresses have a `addr_no` but for some authors this information is missing. In the case where the authors of a publication does not have a `addr_no` and the publication only have one address, it is assumed that all authors on the publication share this address.

3.8 Counting citations

For each reference in the reference list of a publication Clarivate Analytics tries to match the item to one of the records in the database. If there is a match the UID of the cited publication is added to the reference information. In addition to this, each reference can contain information about for example the first author and which year the cited work is published.[4] For publications published in 1982 44% of all references were made to publications outside the database. Since then this number have decreased steadily and in 2014 25% of all references were made to publications outside the database.

In some scientific areas there are citing traditions leading to a publication to occur several times in the same reference list. The reference may for instance point to different pages of the referred publication. In the SRC database these duplicate references are only being counted once.

There are cases when a publication is citing a publication that has a later publication year than itself. In theory this should be impossible and in some cases this is because the algorithm for matching references did not work properly. For example when the title of two publications and the name of their authors are very similar it can happen that the algorithm chooses the wrong publication. But because of different publishing routines there can be legitimate cases when this occur. Therefore the SRC includes citations to publications published one year after the citing publication. That is, if a publication published in 2002 is citing a publication published in 2003 the citation gets counted but if it cites a publication published in 2004, the citation does not get counted.

Citations to publications can be counted using different time windows. The SRC uses three types of time windows: a 3-year window, a 6-year window and an open window. A 3-year window means that when counting the number of citations to a publication, we only count citations made by publications published the same year, or up to two years after the publication i question. That is, for a publication published in 2002 we count citations from publications published 2002-2004. The 6-year window is defined in the same way and using an open window we count all citations to a publication, regardless of when it was made.

References between publications are usually considered to reflect some kind of scientific recognition and the number of citations that a publication receive can thus be said to be a measure of the amount of scientific recognition it has gained. But if researchers refere to their own previous work it is not a measure of recognition from the rest of the research community. Therefore the SRC almost always excludes these self citations when counting the number of citations. The method for removing self citations is based on all author names in both the referring and the cited publication.

[4] The `<reference>` element can contain the following child elements: `UID`, `citedAuthor`, `assignee`, `year`, `page`, `volume`, `citedTitle`, `citedWork`, `doi`, `art_no`, `patent_no`.

If any of the author names (lastname + initials) in the author list of the referring publication is found in the author list of the cited publication, the citation is considered to be a self-citation. No attempt to differentiate between different researchers sharing the same name is done and there is no separate rule for publications with long author lists.

4 Citation based reference values

This section describes the calculation of the reference values that are used for the calculation of the citation based indicators described in Section 5.

4.1 Reference values for subject fields

The *Field reference value* (μ_f) is calculated for each combination of publication year, document type and subject area and is the average number of citations for a publication in each such combination. Given $i = 1, \ldots, P$ publications in a certain subject area, published a certain year of the same document type the field reference value is calculated according to:

$$\mu_f = \frac{\sum_{i=1}^{P} \frac{C_i}{S_i}}{\sum_{i=1}^{P} \frac{1}{S_i}},$$

where

P = the number of publications of the studied document type, the studied year classified as belonging to the subject field in question,
C_i = the number of citations to publication i,
S_i = the number of subject fields publication i has been classified as belonging to.

Publications without any subject field classification are obviously excluded from the calculation, as well as publications without any author address since the latter cannot be part of any country or organisation analysis.

4.2 Reference values for journals

Sometimes it can be of interest to study how much a publication has been cited in relation to other publications in the same journal (and of the same document type and publication year), an indicator called *Journal normalised citation rate*. To be able to do this, a mean citation value for each journal, document type and year has to be calculated. This value is called the *Journal reference value* (μ_j) and is calculated for each journal, identified by the `abbrev_11` title, as:

$$\mu_j = \frac{\sum_{i=1}^{P} C_i}{P}$$

where

P = the number of publications of the studied document type, the studied year published in the journal in question,
C_i = the number of citations to publication i.

Publications without any author addresses are excluded from the calculation of the journal reference value since these cannot be a part of any country or organisation analysis and should therefore not be a part of the reference value for the journal.

4.3 Percentile thresholds

As a complement to the study of publication citation rates in relation to mean values it can be of interest to study the share of publications, for say a university, that are cited more than a specified percentile threshold in a subject field. Commonly used percentile thresholds are 90%, 95% and 99% which indicate that a publication is among the 10%, 5% or 1% most cited in a field if it has yielded more citations than the corresponding percentile threshold value.

The percentile threshold values are calculated for each combination of publication year, subject area and document type, we call such a combination a *field*. Every publication is fractionalised, or weighted, based on the number of subject areas it belongs to. So if a publication is classified as belonging to three subject areas it will be divided into three parts where each part has the weight 1/3. Consider a field with n publications and let $C_1, C_2, \ldots, C_n$ be the number of citations for these publications, ordered from the smallest to the biggest, and let $v_1, v_2, \ldots, v_n$ be the weights associated with these publications. Since it often is the case that multiple publications have the same number of citations the weights for each distinct citation value are added. Let $x_1, x_2, \ldots, x_m$ be the ordered distinct citation values and let $w_1, w_2, \ldots, w_m$ be the corresponding aggregated weights. Then we get the citation value corresponding to the i:th percentile according to:

$$\begin{cases} \frac{1}{2}(x_i + x_{i+1}) & \text{if } \sum_{j=1}^{i} w_j = pW \\ x_{i+1} & \text{if } \sum_{j=1}^{i} w_j < pW < \sum_{j=1}^{i+1} w_j, \end{cases}$$

where $W = \sum_{i}^{m} w_i$ is the sum of the weights for all publications in the field and $p = i/100$ is the desiered percentile in decimal form. That is, pW is the number of weighted publications that is below the i:th percentile.

Since the number of citations to a publication is an integer there will often be many publications that has the same number of citations as the threshold for a certain percentile. In the old database the SRC only considered publications with more citations than the threshold value to be counted as belonging to the corresponding percentile. In practice this meant that, for example, the share of publications among the 10% most cited would be lower than 10%. With the new version of the database the SRC has changed this so that the share of 10% highest cited publications in a field is 10%. This is done by calculating the difference between the number of publications that correspond to 10% of the field and the number of publications that are cited more than the threshold and divide the difference by the number of publications that has the same number of citations as the threshold. This gives us the share of the publications that has the same number of citations as the threshold that should be counted among the 10% highest cited. A simple example illustrates.

Example 1. *The table below show the data for a fictitious field for which we will calculate the threshold for the 90:th percentile.*

i	x_i	w_i	$\sum_{j=1}^{i} w_i$
1	0	521.7	521.7
2	0.5	0.5	522.2
3	1	281.8	804.0
4	1.75	3.0	807.0
5	2	196.7	1003.7
6	3	205.0	1208.7
7	4	161.3	1370.0
8	5	129.7	1499.7
9	6	165.3	1665.0
10	7	117.0	1782.0
11	8	93.3	1875.4
12	9	105.5	1980.9
13	10	104.7	2085.5
14	11	80.7	2166.2
15	41	12.5	2178.7
16	71	3.3	2182.0
17	72	4.8	2186.9
18	73	3.7	2190.5
19	74	2.0	2192.5

In this case $pW = \frac{90}{100} \cdot \sum_{i=1}^{m} w_i = 0.9 \cdot 2192.5 = 1973.3$ *and if we look in the fourth column we see that this value is between 1875.4 and 1980.9 which means that the citation value corresponding to the 90:th percentile is* $x_{12} = 9$.

To get the share of publications that have 9 citations that should be added to the 10% most cited we calculate

$$\frac{0.1 \cdot W - \sum_{j=13}^{19} w_j}{w_{12}} = \frac{219.3 - 211.7}{105.5} = 0.072.$$

That is, of the 105.5 publications that have 9 citations 7.2% (or 7.6 publications) should be among the 10% most cited.

5 Indicators

This section describes how the most commonly used bibliometric indicators are calculated at the SRC. All indicators can be calculated with or without self-citations and in the majority of the cases the SRC will exclude self-citations from the calculations.

5.1 Field normalised citation rate

The field normalised citation rate relates the number of citations to a publication to the average citation rate of a group of comparable publications of the same *document type*, *publication year* and *scientific field*[5].

The SRC calculates the field normalised citation rate (c_f) indicator using a publication fraction oriented method, which means that the number of citations of each subject-address fraction of a publication is normalised against an average citation rate for the same document type, publication year and subject field as the fraction in question belongs to.

When the final average of the normalised citation rate for the analysed unit's publications is calculated, each publication citation rate is weighted by its share of all subject-address fractions for that publication, so that the resulting average will

[5] As defined by the classification of the journal issue made by Clarivate Analytics, after the reclassification of multidisciplinary journals mentioned earlier.

be a weighted average. The SRC average c_f is calculated according to the following formula:

$$c_f = \frac{\sum_{i=1}^{R} \frac{C_i}{S_i \cdot A_i \cdot \mu_{f(i)}}}{\sum_{i=1}^{R} \frac{1}{S_i \cdot A_i}} \tag{1}$$

where

c_f - the average field normalised citation rate,
C_i - the number of citations to the publication of fraction i,
$\mu_{f(i)}$ - the field referece value for the field of fraction i,
R - the number of publication fractions attributed to the analysed unit,
S_i - the number of subject fields the publication of fraction i has been classified as belonging to,
A_i - the total number of author addresses on the publication of fraction i.

5.2 Share of publications on or above a percentile threshold

This indicator is calculated as the share of a unit's publications that have the same number of citations or more than some percentile threshold. For each publication fraction from the unit one of the following is true:

- The publication have fewer citations than the percentile threshold for the subject that the fraction is classified in. In this case the weight of the fraction is not counted at all.
- The publication have the same number of citations as the percentile threshold for the subject that the fraction is classified in. In this case the weight of the fraction gets partially counted by being multiplied by the weight described in Section 4.3.
- The publication have more citations than the percentile threshold for the subject the fraction is classified in. In this case the whole weight of the fraction is counted.

The sum of the weights described above is divided by the sum of all the fractions from the unit to get the unit's share of publications above the percentile threshold in question.

$$p_f = \frac{\sum_{i=1}^{R} \frac{W_{f(i)}}{S_i \cdot A_i}}{\sum_{i=1}^{R} \frac{1}{S_i \cdot A_i}} \tag{2}$$

where

p_f - the share of publications cited on or above a specific percentile for the analysed unit,
R - the number of publication fractions attributed to the analysed unit,
$W_{f(i)}$ - the percentile weight of fraction i. 0 if the number of citations are below the threshold, the weight described in Section 4.3 if the number of citations are the same as the threshold or 1 if the number of citations are larger than the threshold,
A_i - the total number of author addresses on the publication of fraction i,
S_i - the number of subject fields the publication of fraction i has been classified as belonging to.

5.3 Journal normalised citation rate

This indicator shows how an analysed unit's publications are cited in relation to the average citation rate for publications of the same *document type* and *publication year* in the same *journal*.

Since no normalisation against subject fields is performed here, the data set for calculation is only fractionalised on address shares. The indicator is calculated according to the following formula:

$$c_j = \frac{\sum_{i=1}^{R} \frac{C_i}{A_i \cdot \mu_{j(i)}}}{\sum_{i=1}^{R} \frac{1}{A_i}} \tag{3}$$

where

c_j - the average journal normalised citation rate,
R - the number of address fractions attributed to the analysed unit,
C_i - the number of citations to the publication of fraction i,
A_i - the total number of addresses on the publication of fraction i,
$\mu_{j(i)}$ - the journal reference value for the publication of fraction i.

5.4 Journals' field normalised citation rate

Sometimes, it can be of interest to study the average citation rate of the journals a unit publishes in, in relation to the average citation rate of the fields the journal is classified as belonging to.

This indicator compares the average citation rate of the journal's publications of a certain type, a certain year to the average citation rates for publications of the same type and year for the fields the journal is classified in. If the journal is classified in several fields, an average between the normalisation against the fields is calculated. The indicator is calculated according to the following formula:

$$j_f = \frac{\sum_{i=1}^{R} \frac{\mu_{j(i)}}{S_i \cdot A_i \cdot \mu_{f(i)}}}{\sum_{i=1}^{R} \frac{1}{S_i \cdot A_i}} \tag{4}$$

where

j_f - the journal to field normalised citation rate,
R - the number of publication fractions attributed to the analysed unit,
$\mu_{j(i)}$ - the journal reference value for the publication of fraction i,
$\mu_{f(i)}$ - the field reference value for the field where fraction i is classified,
S_i - the number of subject fields the journal of fraction i has been classified as belonging to,
A_i - the total number of author addresses on the publication of fraction i.

5.5 Field normalised share of international publications

This indicator relates a unit's share of international publications to the world average for publications in the same subject area, from the same year, of the same publication type. The calculation mimics the calculation of the field normalised citation rate and is performed as follows:

1. We start by noting if a publication is national or international and denote this with

$$I_{\{0,1\}} = \begin{cases} 0 & \text{for national publications} \\ 1 & \text{for international publications.} \end{cases}$$

   A publication is marked as international if it has at least two different countries among the author addresses.

2. Calculate the expected share of international publications for each subject. Each publication is divided into the same number of fractions as the number of subjects it is classified as belonging to. Suppose there are n fractions in one particular subject area. Then the expected share of international publications for this subject area are calculated according to

$$E_I = \frac{\sum_{i=1}^{n} \frac{I_{\{0,1\}}}{S_i}}{\sum_{i=1}^{n} \frac{1}{S_i}},$$

   where S is the number of subjects that fraction i is classified as belonging to. It is obvious that $0 \leq E_I \leq 1$.

3. We then fractionalise the publications based on the number of addresses and subject areas. Each publication is divided into as many parts as the product of the number of author addresses and the number of subject areas it is classified as belonging to. A publication that are classified as belonging to S subjects and have A author addresses is divided into $S \cdot A$ shares where each share have the weight

$$\frac{1}{S \cdot A}.$$

4. Calculate the field normalised share of international publications for a specific unit. Suppose this unit have $i = 1, \ldots, n$ fractions of which m comes from international publications. Then the field normalised share of international publications are calculated according to

$$\frac{\sum_{i=1}^{m} \frac{1}{E_{I_i}} \cdot \frac{1}{S_i \cdot A_i}}{\sum_{i=1}^{n} \frac{1}{S_i \cdot A_i}},$$

   where $m \leq n$ and $0 < E_I \leq 1$.

The result will be an indicator varying around 1, where a value above 1 indicates that the unit's share of international publications are larger than average and a value below 1 indicates that the unit's share of international publications are smaller than average.

5.6 Field normalised share of non-cited publications

The analysed unit's shares of non-cited publications are compared to the world share of non-cited publications for each combination of subject area, publication year and document type in the same manner as described under *Field normalised share of international publications* above.

The result will be an indicator varying around 1, where a value above 1 indicates that the unit's share of non-cited publications are larger than for the world average and vice versa.

5.7 Simulated stability intervals

Sometimes it can be of interest to see how much an indicator depends on a few extreme values. To get an idea of this we can simulate stability intervals for the indicator in question and this is done in the following way. Assume that the organization in the study has n publications. From these publications a sample of size n is drawn with replacement. This is repeated 1000 times which gives us 1000 samples. We then calculate the indicator in question for each of the 1000 samples and order the results from the smallest to the largest. From this ordered set we then take the 2.5th and the 97.5th percentile and call them the lower and upper bound. See [3] for a more thorough description of stability intervals.

References

[1] Gunnarsson, Fröberg, Jacobsson, Karlsson (2011) *Subject classification of publications in the ISI database based on references and citations*, www.vr.se.

[2] Sjöstedt, Aldberg, Jacobsson (2014) *Guidelines for using bibliometrics at the Swedish Research Council*, www.vr.se.

[3] Waltman et. al. (2012) *The Leiden Ranking 2011/2012: Data Collection, Indicators, and Interpretation*, Journal of the American Society for Information Science and Technology, Vol. 63, Issue 12, p. 2429.

A Document types

The table below show the share of publications for each document type as classified by Clarivate Analytics and the SRC respectively. The document type *Article* which all SRC analyses are based on constitutes 63% of all publications.

	Share of publications	
Document type	Clarivate (%)	SRC (%)
Article	59	63
Meeting Abstract	11	11
Proceedings Paper	10	10
Book Review	6	6
Editorial Material	4	4
Letter	3	3
Review	3	0
Note (coded as Article effective 1996)	2	0
News Item	1	1
Poetry	0	0
Correction	0	0
Biographical-Item	0	0
Art Exhibit Review	0	0
Correction, Addition	0	0
Item About An Individual	0	0
Record Review	0	0
Film Review	0	0
Music Performance Review	0	0
Fiction, Creative Prose	0	0
Discussion (coded as Editorial effective 1996)	0	0
Theater Review	0	0
Dance Performance Review	0	0
Reprint	0	0
Software Review	0	0
Music Score Review	0	0
Bibliography	0	0
TV Review, Radio Review	0	0
Excerpt	0	0
TV Review, Radio Review, Video	0	0
Hardware Review	0	0
Script	0	0
Database Review	0	0
Chronology (coded as Article effective 1996)	0	0
Music Score	0	0
Main Cite	0	0
Meeting Summary	0	0
Abstract of Published Item	0	0
Book	0	0