# Formal definitions of field normalized citation indicators and their implementation at KTH Royal Institute of Technology

Per Ahlgren and Peter Sjögårde, 2015-02-17

## Introduction

This document describes the calculation of bibliometric indicators based on field normalization in the bibliometric database at KTH (Bibmet), which is based on Web of Science data. The indicators are described in Part 1 and aspects regarding implementation in the KTH database are addressed in Part 2.

The following indicators are defined in this document:

- mean field normalized citation rate (cf)
- top10% publications (ptop10%)
- mean field normalized journal impact (jcf)
- proportion publications in the 20% most frequently cited journals in the field (jtop20%)

## Part 1 Definitions

This document treats the case, in which fractional counts are used in the calculations of indicator values. In case whole counts should be used in the calculations, $a_i$ in Eq. (1) below is set to unity.

Let A be a unit of analysis, and *n* the number of publications for A. Let $r_i$ be the number of authors of the *i*th publication for A. Let $a_i$ be the *fraction* A has of the *i*th publication. We consider two cases.

(**1**) A is an organization. We treat two subcases. (**1.1**) $a_i$ is the *author* fraction A has of the *i*th publication and is defined as

$$a_i = \sum_{j=1}^{m_i} \frac{1}{r_i s_j} \qquad (1)$$

where $m_i$ is the number of authors affiliated to A regarding the *i*th publication, and $s_j$ the number of affiliations of the *j*th of these A authors. Note that the right-hand side in Eq. (1) is equal to $m_i/r_i$ when each A author has exactly one affiliation. (**1.2**) $a_i$ is the *organization* fraction A has of the *i*th publication and is defined as the number of occurrences of A's name in the address field of the *i*th publication divided by the total number organization name occurrences in the address field in question.

(**2**) A is an individual author. $a_i$ is the *author* fraction A has of the *i*th publication and is in this case defined as $1/r_i$.

### 1.1 Mean field normalized citation rate

We define the *mean field normalized citation rate* for A, *mcf*(A), as

$$mcf(\mathrm{A}) = \frac{\sum_{i=1}^{n} a_i \overline{x}_i}{\sum_{i=1}^{n} a_i} \tag{2}$$

$$\overline{x}_i = \left(1/q_i\right) \sum_{q=1}^{q_i} c_i / \mu_{iq}$$

$$\mu_{iq} = \frac{\sum_{j=1}^{m_{iq}} c_j / F_j}{\sum_{j=1}^{m_{iq}} 1/F_j}$$

where $q_i$ ($c_i$) is the number of subject categories (the citation rate) of the *i*th publication for A, $m_{iq}$ is the number of publications, with the same publication year and of the same document type as the *i*th publication for A, in the *q*th subject category of the *i*th publication of A, and $c_j$ ($F_j$) the citation rate of the *j*th of these publications (the number of subject categories of the *j*th of these publications). $\mu_{iq}$ is the *field reference value* that the citation rate of the *i*th publication, $c_i$, is normalized against regarding the *q*th subject category of the publication, and the normalization gives rise to a *field normalized citation rate* for the publication.

### 1.2. ptop10%

We define *ptop10%* for A, *ptop10%*(A), as

$$ptop10\%(\mathrm{A}) = \frac{\sum_{i=1}^{n} a_i \sum_{q=1}^{q_i} b_{iq}}{\sum_{i=1}^{n} a_i} \tag{3}$$

$$b_{iq} = \frac{1}{q_i} \times \frac{\max(y_{iq}^{c_i+1} - \max(0.9, y_{iq}^{c_i}), 0)}{y_{iq}^{c_i+1} - y_{iq}^{c_i}}$$

where $q_i$ ($c_i$) is the number of subject categories (the citation rate) of the *i*th publication for A, $y_{iq}^{c_i}$ ( $y_{iq}^{c_i+1}$ ) the proportion publications–with respect to the citation distribution, which concerns publications with the same publication year and of the same document type as the *i*th publication for A, and belonging to the *q*th subject category of this publication–with less than $c_i$ ($c_i$ + 1) citations.[1]

[1] Weights are used for the citation distributions at stake. Each citation value in a given distribution is assigned the weight 1/*k*, where *k* is the number of subject categories of the corresponding publication. The weight is the fraction with which the publication contributes to each of its subject categories. The proportion publications with less than *c* citations is then the sum of the weights for the citation values that are less than *c*, divided by the sum of weights for all the citation values in the distribution.

$\max(y_{iq}^{c_i+1} - \max(0.9, y_{iq}^{c_i}), 0) \,/\, y_{iq}^{c_i+1} - y_{iq}^{c_i}$ is the fraction of the *i*th publication with which the publication is assigned to the 10% most cited publications. Observe that this fraction is weighted by $1/q_i$, i.e., by the fraction of the publication that belongs to the *q*th subject category. The approach to assign fractions of publications to the (for instance) 10% most cited publications is described and discussed by Waltman and Schreiber (2013).

### 1.3 Mean field normalized journal impact

We define the *mean field normalized journal impact* for A, *mjcf*(A), as

$$mjcf(\mathrm{A}) = \frac{\sum_{i=1}^{n} a_i jcf_i}{\sum_{i=1}^{n} a_i} \tag{4}$$

$$jcf_i = \frac{\sum_{j=1}^{p_i} \overline{x}_j}{p_i}$$

$$\overline{x}_j = \left(1/F_{ij}\right) \sum_{q=1}^{F_{ij}} c_j / \mu_{jq}$$

$$\mu_{jq} = \frac{\sum_{k=1}^{m_{jq}} c_k / F_k}{\sum_{k=1}^{m_{jq}} 1/F_k}$$

where $jcf_i$ is the *mean field normalized citation rate of the journal*, say $J_i$, *of the* *i*th *publication*, $p_i$ the number of publications in $J_i$, $c_j$ the citation rate of the *j*th publication in $J_i$, say $P_j$, $F_{ij}$ the number of subject categories of $P_j$, $m_{jq}$ the number of publications, with the same publication year and of the same document type as $P_j$, in the *q*th subject category of $P_j$, and $c_k$ ($F_k$) the citation rate (the number of subject categories) of the *k*th of these publications. $\mu_{jq}$ is the field reference value that the citation rate of $P_j$, $c_j$, is normalized against regarding the *q*th subject category of $P_j$, and the normalization gives rise to a field normalized citation rate for $P_j$ (cf. the definition of mean field normalized citation rate above). (If $J_i$ is a non-multidisciplinary journal, $F_{ij} = F_{i(j+1)}$ ($j = 0, 1, \ldots, p_i - 1$), since the number of subject categories of a publication in $J_i$ is then equal to the number of subject categories of $J_i$.)[2]

### 1.4 Proportion publications in the 20% most frequently cited journals in the field

We define the *proportion publications in the 20% most frequently cited journals in the field* for A, *jtop20%*(A), as

$$jtop20\%(\mathrm{A}) = \frac{\sum_{i=1}^{n} a_i \sum_{q=1}^{F_i} b'_{iq}}{\sum_{i=1}^{n} a_i} \tag{5}$$

[2] Cf. Section 2.5 below.

$$b'_{iq} = \frac{1}{F_i} \times \frac{\max(\min(0.2, y_{iq}^{r_{iq}}) - y_{iq}^{r_{iq}-1}, 0)}{y_{iq}^{r_{iq}} - y_{iq}^{r_{iq}-1}}$$

where $F_i$ is the number of subject categories of the journal, say $J_i$, of the $i$th publication of A, $r_{iq}$ the rank of $J_i$ in the ranking of the journals in the $q$th subject category of $J_i$, where the journals are ranked descending after their mean field normalized citation rates[3], $y_{iq}^{r_{iq}}$ ($y_{iq}^{r_{iq}-1}$) the proportion publications appearing in the journals–regarding the ranking of the journals in the $q$th subject category of $J_i$–with a rank less than or equal to $r_{iq}$ ($r_{iq}$–1).[4] The rightmost factor in $b'_{iq}$ is the fraction of $J_i$ with which $J_i$ is assigned to the 20% most frequently cited journals in the $q$th subject category of $J_i$. Observe that this fraction is weighted by $1/F_i$, i.e., by the fraction of $J_i$ that belongs to the $q$th subject category. The approach to assign fractions of journals to the (for instance) 20% most cited journals is basically the same as the assignment approach used in the definition of *ptop10%* (Eq. (3)).

## Part 2 Implementation at KTH Royal Institute of Technology

### 2.1 Database contents

The bibliometric database at KTH (Bibmet) contains the following indexes:

- Science Citation Index Expended (SCIE)
- Social Sciences Citation Index (SSCI)
- Arts & Humanities Citation Index (AHCI)
- Conference Proceedings Citation Index - Sciences (CPCI-S)
- Conference Proceedings Citation Index - Social Sciences & Humanities (CPCI -SSH)).

SCIE, SSCI, AHCI from 1980 and CPCI-S and CPCI-SSH from 1990.

### 2.2 Document types included in calculations

In Bibmet, calculations are made for all combinations of document types, publication years and Web of Science categories. However, the default presentation of field normalized citation indicators concern only articles and reviews. The reason for excluding other document types is the risk for anomalies caused by a low number of publications in the reference groups and question marks regarding data quality and citation matching (this especially applies to proceedings papers).

### 2.3 Citations included

[3] Note that for a given journal in the ranking, these rates may vary across the different rankings, corresponding to different subject categories, in which the journal occurs.

[4] A non-multidisciplinary journal in the ranking contributes, if each of its publications has a field reference value, with respect to the $q$th subject category of $J_i$, greater than or equal to 0.5 (see Section 2 below), with $(1/k)m$ publications to the ranking, where $k$ is the number of subject categories of the journal and $m$ the number of publications of the journal.

Citations from all records in the database are included in the calculations. The indicators are calculated both with self-citations included and excluded. The default presentation is made with self-citations excluded, since the intention when calculating citation indicators is to see what impact a publication has had on other researchers than those who wrote the publication. Furthermore, one should avoid giving incentives to systematic self-quotation.

## 2.4 Retroactive changes of the Web of Science subject category assigned to journals

If a journal is reclassified from one Web of Science subject category to another by Thomson Reuters (TR), no retroactive changes are made in the delivered raw data. However, in Web of Science TR changes the classification retroactively. Changes of the classification affect the field reference values and consequently the outcome of the calculations described in this document. For Bibmet to be consistent with Web of Science, retroactive changes of the Web of Science subject categories assigned to journals are made in Bibmet.

## 2.5 Reclassification of journals categorized as Multidisciplinary in Web of Science

The large (in terms of publications output) and highly prestigious journals Nature, Science and PNAS are classified by TR as multidisciplinary. When field normalization is applied the classification of these highly cited journals into the same category results in very high field reference values for this "field". By reclassifying publications in journals within the multidisciplinary subject category according to their "real" topics the publications are instead compared to other publications within the same subject field. The Swedish Research Council has developed and applied a methodology for reclassification of publications within the multidisciplinary Web of Science subject category into other categories based on citations (Gunnarsson, Fröberg, Jacobsson, & Karlsson, 2011). It enables a higher degree of like-to-like comparison. The same methodology is used at KTH.

## 2.6 Exclusion of publication fractions with low field reference values

For all the four indicators defined in Part 1, publication fractions with field reference values less than 0.5 are excluded.[5]

**Example 2.1** (*mcf* and *ptop10%*). Assume that the *i*th publication of A, say $P_i$, belongs to three subject categories and that exactly one of these categories has a field reference value less than 0.5 (regarding publications with the same publication year and of the same document type as $P_i$). For Eq. (2), $a_i$ in the denominator and $\overline{x}_i$ in the numerator are then multiplied by 2/3, and $q_i$ is equal to 2 (and not to 3). Thus, the sum of $\overline{x}_i$ concerns two field normalized citation rates for $P_i$, and the sum is multiplied by 1/2 (and not by 1/3).

[5] Such field reference values might give rise to very large field normalized citation rates in spite of few citations.

For Eq. (3), under the assumptions given, $a_i$ in the denominator and the rightmost sum in the numerator are multiplied by 2/3, and $q_i$ is equal to 2 (and not to 3). Thus, the sum concerns two ratios, both of which are weighted by 1/2 (and not by 1/3). ■

**Example 2.2** (*mjcf* and *jtop20%*). Assume that the journal, $J_i$, of the *i*th A publication belongs to four subject categories. Assume that for the *j*th publication in $J_i$, say $P_j$, published a given year and of a given document type, two of the four subject categories have a field reference value less than 0.5. For Eq. (4), 2/4 is subtracted from the denominator of $jcf_i$, and $\overline{x}_j$ in its numerator is multiplied by 2/4. $F_{ij}$ is equal to 2 (and not to 4). Thus, the sum of $\overline{x}_j$ concerns two field normalized citation rates for $P_j$, and the sum is multiplied by 1/2 (and not by 1/4).

For *jtop-20%* (Eq. (5)), under the assumptions given, two of the four rankings in which $J_i$ occurs are such that $P_j$ does not contribute to the mean field normalized citation rates of $J_i$ in the rankings. ■

## References


Gunnarsson, M., Fröberg, J., Jacobsson, C., & Karlsson, S. (2011). *Subject classification of publications in the ISI database based on references and citations* (No. 4).

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology, 64*(2), 372-379.