Scientometrics (2019) 120:283–300
https://doi.org/10.1007/s11192-019-03121-z

Comparison of publication-level approaches to ex-post citation normalization

Cristian Colliander[1,2] · **Per Ahlgren**[3,4]

Received: 7 January 2019 / Published online: 17 May 2019

Abstract

In this paper, we compare two sophisticated publication-level approaches to ex-post citation normalization: an item-oriented approach and an approach falling under the general algorithmically constructed classification system approach. Using articles published in core journals in Web of Science (SCIE, SSCI & A&HCI) during 2009 ($n = 955{,}639$), we first examine, using the measure Proportion explained variation (PEV), to what extent the publication-level approaches can explain and correct for variation in the citation distribution that stems from subject matter heterogeneity. We then, for the subset of articles from life science and biomedicine ($n = 456{,}045$), gauge the fairness of the normalization approaches with respect to their ability to identify highly cited articles when subject area is factored out. This is done by utilizing information from publication-level MeSH classifications to create high quality subject matter baselines and by using the measure Deviations from expectations (DE). The results show that the item-oriented approach had the best performance regarding PEV. For DE, only the most fine-grained clustering solution could compete with the item-oriented approach. However, the item-oriented approach performed better when cited references were heavily weighted in the similarity calculations.

Keywords Algorithmically constructed classification system approach · Citation impact · Field normalization · Item-oriented approach · Research evaluation

✉ Cristian Colliander
cristian.colliander@umu.se

Per Ahlgren
per.ahlgren@uadm.uu.se

[1] Department of Sociology, Inforsk, Umeå University, Umeå, Sweden

[2] University Library, Umeå University, Umeå, Sweden

[3] Department of Statistics, Uppsala University, Uppsala, Sweden

[4] KTH Library, KTH Royal Institute of Technology, Stockholm, Sweden

Introduction

Over the past two decades, bibliometric indicators based on citations have become increasingly more important in research assessments of scientific influence. Such assessments are sometimes multidisciplinary assessments, in which publications from different research fields are compared. Especially in these multidisciplinary cases, some form of field-normalization of citations is called for. There are subjects, and thereby the fields to which they are associated, that attract a lot of citations from adjacent subjects, tend to refer to recent publications and have long reference lists of which a large proportion of the cited references point to publications in the database. Such subjects will on average receive more citations per publication than subjects that attract few citations from adjacent subjects, tend to refer to older publications and have short reference lists of which many cited references point to publications outside the database. The indicated factors are not the only ones involved, but they undoubtedly have an effect on citation counts.

One can distinguish between two general approaches for removing or decreasing the variation in the observed distribution of citation counts over publications that arises from the publications' disparate subject matter: ex-post and ex-ante normalization. Ex-post citation normalization tries to adjust a raw enumerated citation count by relating it to a reference value calculated from a set of similar publications (e.g., Braun and Glänzel 1990; Colliander 2015; Moed et al. 1995; van Raan 1996; Waltman et al. 2011a, b), whereas ex-ante citation normalization tries to arrive at a normalized citation score in the enumeration stage by some form of fractional counting (e.g., Glänzel et al. 2011; Leydesdorff and Bornmann 2011; Zitt and Small 2008; Zitt 2010, 2013). Although both methods have their pros and cons (Glänzel and Moed 2013), in this work we deal exclusively with ex-post normalization.

In ex-post normalization, journal-based approaches have been–and still are–the norm. Typically, the reference sets for the target publications (i.e. the publications subject to normalization) are obtained from the journal subject categories in the Web of Science (WoS) database (Clarivate Analytics): the field of a target publication is defined as the WoS subject category, or categories, to which the journal of the publication has been (manually) classified. However, there are drawbacks with these journal-based approaches. For instance, it is possible that the subfields of a certain field, where the fields are defined at a given level of granularity, differ substantially from each other in terms of citation density (e.g., Adams et al. 2008; Neuhaus and Daniel 2009; van Eck et al. 2013; van Leeuwen and Medina 2012; Zitt et al. 2005). Thus, using journal-based approaches in assessments of scientific influence might give rise to instances of the comparing-apples-with-oranges phenomenon.

Nowadays, though, there exist publication-level approaches to ex-post citation normalization. The CWTS Leiden Ranking 2018[1] involves a publication-level classification system obtained by use of a methodology proposed by Waltman and van Eck (2012). WoS publications of the types *Article* and *Review* are clustered on the basis of direct citation links between them, and the clustering technique used is similar to modularity-based clustering (Newman 2004a, b). Generated clusters (4047 so-called micro-level fields) are then used for the identification of reference sets for the target publications of the ranking. Quite recently, Colliander (2015) proposed an item-oriented approach to ex-post normalization.

[1] http://www.leidenranking.com/ranking/2018/list.

In this approach, for a given item, in this context a given target publication, a reference set of thematically similar publications is automatically identified—without any clustering—and used as the basis for deriving a reference value. The publication–publication similarity estimation make use of both textual and citation-based information.

A number of earlier studies have dealt with the question how to best correct for variation in the observed citation distribution that stems from subject matter heterogeneity. Leydesdorff et al. (2013) and Radicchi and Castellano (2012) compared ex-post normalization approaches to ex-ante approaches. Four normalization approaches, one ex-post and three ex-ante, were compared by Waltman and Eck (2013a) in a large-scale study. This study used an improved evaluation methodology relative to the two works referred to above: *different* classification systems were used in the implementation and the evaluation of the ex-post normalization approach in order not to give this approach an advantage over the other approaches (Sirtes 2012). Regarding what it means to correct differences in citation practices between fields, Waltman and Eck (2013a) used the following idea: the degree to which differences in citation practices between fields have been corrected is indicated by the degree to which the field-normalized citation distributions coincide with each other (Crespo et al. 2013). This idea has also been utilized in works where the effectiveness of ex-post normalization approaches has been compared (e.g., Li et al. 2013; Li and Ruiz-Castillo 2013; Perianes-Rodriguez and Ruiz-Castillo 2017).

In this study, we compare two sophisticated publication-level approaches to ex-post citation normalization: an approach that is similar to the approach used in the CWTS Leiden Ranking 2018, the latter approach briefly described above, and an approach similar to the one proposed by Colliander (2015). We also include, as a benchmark, a traditional journal-based approach that is based on the WoS journal subject categories.

The remainder of this paper is organized as follows. In the next section, we describe the dataset of the study. In the third section, the approaches to ex-post citation normalization that we compare are described in detail. In the fourth section, we put forward our findings. In the final section, we discuss the findings and provide conclusions.

Data

The data source of the study was Bibmet, the bibliometric version of WoS at KTH Royal Institute of Technology (Sweden). Bibmet contains more than 50 million WoS publications, with the earliest publication year equal to 1980, and is updated quarterly. The dataset consists of all publications published year 2009, of the document type *Article*, included in the three journal indexes of WoS (SCI-EXPANDED, SSCI and A&HC) and published in journals classified as core journals by the CWTS Leiden Ranking 2015. To be classified as a core journal, a journal should have an international scope and a sufficiently large number of cited references to other core journals suggesting that the journal belongs to a field that is suitable for citation analysis (CWTS Leiden Ranking 2015, Methodology 2015). It turned out that 955,639 publications satisfied the stated conditions.

Ex-post normalization approaches

In an ex-post normalization setting, the *reference value* for the citation count (x_i^c) of a target publication (x_i) can be given by:

$$\mathrm{RF}(x_i) = \frac{\sum_{x_j \in D, x_j \neq x_i} w_{ij} \phi\left(x_j^c\right)}{\sum_j w_{ij}} \tag{1}$$

where D is the set of publications (usually of the same publication type and publication year as x_i) in the bibliographic universe considered and $\phi(\cdot)$ typically the identity function or some function that transform the raw citation count. The weight w_{ij} determines the influence x_j have on the reference value, and this weight is the main difference between ex-post normalization approaches. A normalized version of a publications citation count ($\hat{x}_i^c$) can then, for example, be calculated as

$$\hat{x}_i^c = \phi(x_i^c) \Big/ \mathrm{RF}(x_i) \tag{2}$$

or

$$\hat{x}_i^c = \phi(x_i^c) - \mathrm{RF}(x_i) \tag{3}$$

where (2) and (3) express the difference between the observed citation count and the reference value in relative and absolute terms, respectively. For all calculations in this paper, ϕ is the natural logarithm of the result of adding 1 to the raw citation count (Thelwall 2019). Moreover, the set D of Eq. (1) is in our case identical to the publication dataset of the study. Thus, the publications taken into consideration in calculating a reference value for an article x_i belonging to the dataset are all articles and published in year 2009.

Traditional journal-based approach

We include the de facto standard for normalization of citation counts as a benchmark to the other considered approaches in this work. In that standard, the journal in which the publication is published is used as a proxy for the subject matter of the publication. Here, we use the journal subject categories of WoS. The corresponding classification system allows a journal (and thus its publications) to be classified into multiple categories, so w_{ij} express some notion of the number of shared categories between x_i and x_j contrasted with the total number of categories that x_i and x_j belong to. If we let x_i^g denote the set of categories x_i belongs to then

$$w_{ij} = \frac{\left|x_i^g \cap x_j^g\right|}{\left|x_i^g \cup x_j^g\right|} \tag{4}$$

can be used as an estimation of the similarity between x_i and x_j and is equal to the Jaccard similarity coefficent for x_i^g and x_j^g. We refer to the traditional journal-based approach used in this work as "WoS SC".

Algorithmically constructed classification system approach

A recent trend is to forgo journal-based classification systems and create a subject scheme based on publication-level data and clustering routines. In this work, we use a classification system obtained on the basis of a methodology of Waltman and van Eck (2012), briefly

described in the introduction of this paper. The system, which has been implemented by the bibliometric group at KTH, is hierarchical and has four levels of clusters, where, for each level, the clusters are pairwise disjoint. The version of the system used in this study is 2017 Quarter 3, and this version contains about 28 million Bibmet publications, where each publication is of the document type *Article* or *Review* (note, though, as mentioned above, that our dataset only contains publications of the type *Article*). The clustering solution of the top level (level 4) of the system has 28 clusters, whereas the solution of the next level (3) has 722 clusters. The solution of the next to bottom level (2) has 4268 clusters, and the solution of the bottom level (1) has 35,026 clusters. The different granularities of the four solutions correspond to different values of the resolution parameter involved in the Waltman and van Eck methodology. We use all four clustering solutions for citation normalization in our analysis.

For each cluster in the classification system, regardless of the level of the cluster, labels describing the content of the cluster have been automatically obtained on the basis of author keywords, journal names, names of WoS subject categories and words derived from address data.

The cluster size distributions are right-skewed and the median cluster size (median absolute deviation) for the lowest resolution, i.e., level 4, to the highest resolution, i.e., level 1, are: 31,780 (16,601), 1037.5 (570.5), 149 (114) and 18 (13).

Given a classification system constructed in this way, as it consists of mutual exclusive groups (clusters) for each level, w_{ij} is binary and equals 1 if x_i and x_j belong to the same group, 0 otherwise. Note that in this case, Eq. (1) simplifies to the average over the (possibly transformed) citation counts of the publications in the group of x_i (though not including $\phi(x_i^c)$ in the calculation of this average).

We use four approaches, under the general algorithmically constructed classification system approach, in this work, where these approaches differ from each other only with respect to classification system level. Let "DCC L4" ("DCC" for "Direct Citation Clustering", "L" for "Level"), "DCC L3", "DCC L2" and "DCC L1" denote the four approaches.

Item-oriented approach

An item-oriented approach does not make use of any clustering. Instead it makes use of a more direct methodology to calculate w_{ij}. Any publication-level features can be used in an effort to estimate the subject similarity between x_i and x_j. The most obvious features are those derived from text and the reference lists of the publications. It is further compelling to use these feature sets in combination, i.e., in a hybrid similarity setting.

From the titles and abstracts of the 955,639 articles of the study, we extract nouns and adjectives, which are then stemmed, and from the reference lists we extract cited references that are processed by an automatic method similar to the one described in (Colliander and Ahlgren 2012) to partly correct cases where several distinct references, occurring in different publications, might represent the same publication due to spelling variation. After these operations, we represent each article x_l by two vectors, one corresponding to textual data, one to cited references. For the former vector, we use the tf-idf (term frequency-inverse document frequency) scheme for generating term (i.e. stem) weights. The *weight for article* x_l *with respect to term* q ($\omega_q^{\text{term}}(x_l)$) is defined as

$$\omega_q^{\text{term}}(x_l) = \text{freq}_{ql} \times \log\left(\frac{n}{n_q}\right) \tag{5}$$

where freq_{ql} is the frequency of term q in x_l, n the number of considered articles (=955,639), and n_q the number of considered articles in which q occurs. In the vector corresponding to x_l, the weight occurs in the qth position.

Regarding the vector for x_l that corresponds to cited references, we use the idf part of Eq. (5), i.e. the rightmost factor. The *weight for article x_l with respect to cited reference r* ($\omega_r^{\text{cr}}(x_l)$) is defined as

$$\omega_r^{\text{cr}}(x_l) = a_{rl} \log\left(\frac{n}{n_r}\right) \tag{6}$$

where a_{rl} is 0 or 1, depending on if r is absent or present in x_l, respectively, and n_r the number of considered articles in which r occurs. In the vector corresponding to x_l, the weight occurs in the rth position.

If we let $\text{sim}_{\text{cr}}(x_i, x_j)$ and $\text{sim}_{\text{term}}(x_i, x_j)$ denote the cosine similarity between (the corresponding vectors of) x_i and x_j when using cited references and terms respectively, then:

$$w_{ij} = \gamma \times \widehat{\text{sim}}_{\text{cr}}(x_i, x_j) + (1-\gamma) \times \widehat{\text{sim}}_{\text{term}}(x_i, x_j) \tag{7}$$

where

$$\begin{aligned} \widehat{\text{sim}}_{\text{cr}}(x_i, x_j) &= \frac{\text{sim}_{\text{cr}}(x_i, x_j)}{\sum_{k \neq i} \text{sim}_{\text{cr}}(x_i, x_k)} \\ \widehat{\text{sim}}_{\text{term}}(x_i, x_j) &= \frac{\text{sim}_{\text{term}}(x_i, x_j)}{\sum_{k \neq i} \text{sim}_{\text{term}}(x_i, x_k)} \end{aligned} \tag{8}$$

are local per article normalization of the similarity values. The step corresponding to Eq. (8) is necessary as the similarity value distributions differ depending on if cited references or terms are used. The normalization helps making the γ parameter interpretable, i.e. any deviation from $\gamma = 0.5$ gives unequal weight to the different feature sets in the hybrid similarity value, w_{ij}, given by Eq. (7).

When calculating cosine similarity based on text and cited references, we only consider the k-nearest neighbors (the k articles with the highest cosine similarity values, k being a parameter) for each feature set. If, say, $k = 20$, this means that the number of unique articles influencing the reference value is at maximum 40. This is partly for efficiency reasons (k-nearest neighbors can be calculated efficiently in large data sets) and partly conceptual (the large tail of very small similarity values for each article is of little interest in this context). Note that, regarding the fractions of Eq. (8), the two sim functions, corresponding to cited references and terms, are the sim functions obtained *after* application of the k-nearest neighbor approach. In this setting, $\text{sim}_{\text{cr}}(x_i, x_j)$ can be 0 while $\text{sim}_{\text{cr}}(x_j, x_i) > 0$, or vice versa. The same is of course true for the term-based similarity as well. Indeed, there are exactly four possibilities regarding the two sim values for x_i and x_j, regardless of which type of similarity that is considered: both values are equal to 0, since x_i (x_j) does not belong to the k most similar articles for x_j (x_i); one of the values is equal to the original sim value (x_j belongs to the k most similar articles for x_i or conversly), whereas one of the values is equal to 0; both values are equal to the two original sim values.

Remark

Our analysis involves two publication-level approaches to ex-post citation normalization, as well as a traditional journal-based approach, which uses WoS journal subject categories. Recall that the main difference between these three approaches is how the weight w_{ij} is defined. The corresponding three weight definitions are given by Eqs. (4) and (7), and in the next to last paragraph of the sub section "Algorithmically constructed classification system approach". These equations/paragraph thereby play an important role in this paper.

Results

In this section, we evaluate the relative performance of the considered ex-post citation normalization approaches in two ways. We first examine to what extent the approaches can explain and correct for variation in the citation distribution that stems from subject matter heterogeneity. Then we consider a subset of the articles from life science and biomedicine and evaluate the fairness of the approaches with respect to their ability to identify highly cited articles when subject area is factored out.

Proportional reduction of variation in the citation distribution

We investigate the proportion of variation in the citation distribution that can be explained by taking the subject matter of the articles into consideration. We can consider the derivation of reference values by Eq. (1) as a constrained prediction exercise, where we use information regarding an articles subject matter to predict its citation count.[2] Explained variation can be considered a measure of the relative gains in prediction accuracy when prediction is based on Eq. (1) rather than just using the overall mean as the best guess, which would be equal to no normalization. Thus, we define the proportion of explained variation, PEV, as:

$$\mathrm{PEV} = \frac{\sum_{i=1}^{n}\left(\phi(x_i^c) - \frac{\sum_{j=1}^{n}\phi(x_j^c)}{n}\right)^2 - \sum_{i=1}^{n}\left(\phi(x_i^c) - RF(x_i^c)\right)^2}{\sum_{i=1}^{n}\left(\phi(x_i^c) - \frac{\sum_{j=1}^{n}\phi(x_j^c)}{n}\right)^2} \tag{9}$$

where n is the number of articles.

Figure 1 shows how the different normalization approaches affect PEV.

While the DCC approaches only have one parameter, i.e. the resolution parameter, which influences the granularity of the clustering solutions of the classification system, this

[2] In rare cases, the DCC approaches and the item-oriented approach might not produce a reference value for a target publication. With respect to DCC, this can happen if the publication ends up in a singleton cluster and in the item-oriented case if a publication does not share any references/terms with other publications. Instead of just removing these publications, we penalize such outcome in the evaluation of the approaches by using the overall mean of the citation distribution as the reference value. This should have minimal impact on the results presented here, though, as less than 0.3% of the publications are affected. In an applied setting, however, this might constitute a problem that demands manual ad-hoc fixes.

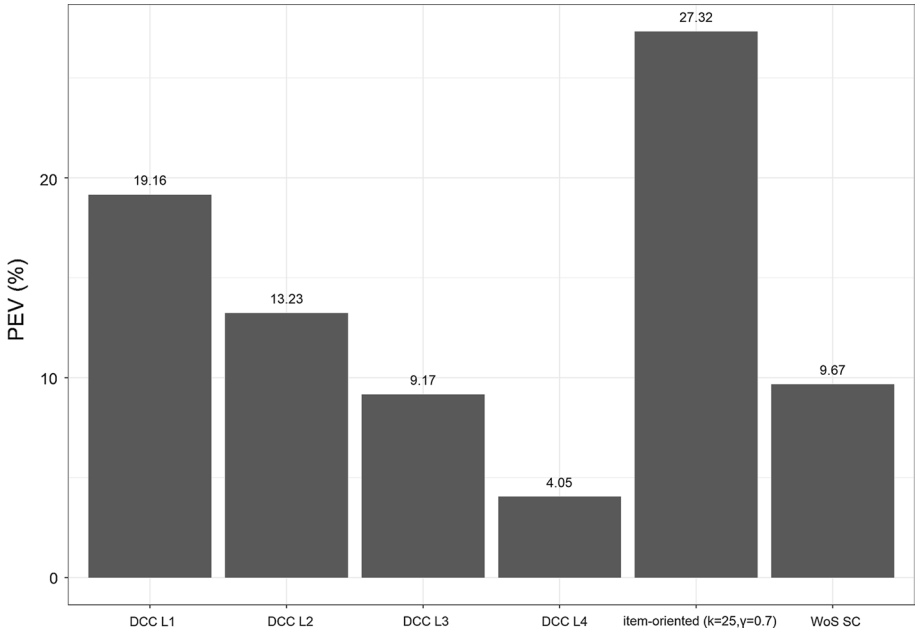

Fig. 1 PEV (as percentages) when using weights in Eq. (1) derived from the approaches DCC L1-L4, the item-oriented approach (the latter with parameters $k = 25$ and $\gamma = 0.7$, see more on this below) and WoS SC

parameter of course has a massive effect on PEV. DDC L4 and DDC L3, corresponding to low resolution clustering solutions, perform worse than WoS SC, the approach based on the traditional journal classification approach (Fig. 1). The item-oriented approach has two parameters whose values are not easily chosen. To gauge the effect of k in k-nearest neighbors and γ in the hybrid similarity calculations, we consider PEV values over this parameter space in Fig. 2. The contour plot of the figure shows how PEV-values relate to the number of k-nearest neighbors and the γ weight and provides a two-dimensional view in which all points that have the same PEV-values are connected to produce contour lines.

As can be seen, the PEV approach has its maximum when cited references are weighted heavily (γ around 0.7) and the number of nearest neighbors for each feature set is relatively low (around 10–30). But the PEV values are pretty stable with respect to these parameters as long as not very low γ weighs are used as then the PEV values drop markedly. Note though, that adding the term-based feature set into the reference value calculation (i.e. $\gamma < 1.0$) in fact reduces PEV in most cases. With $k \approx 25$ and $\gamma \approx 0.7$, the item-oriented approach has the best PEV performance, followed by DCC L1, which corresponds to the most fine-grained clustering solution (Fig. 1; PEV values equal to 27.32 and 19.16, respectively).

Expected proportions of top *z*% articles in subject groupings derived from MeSH

Another way to gauge the effect of the considered citation normalization approaches is to make use of an external subject classification scheme that—for the sake of this exercise at least–can be considered as a "ground truth" with respect to subject matter. Such

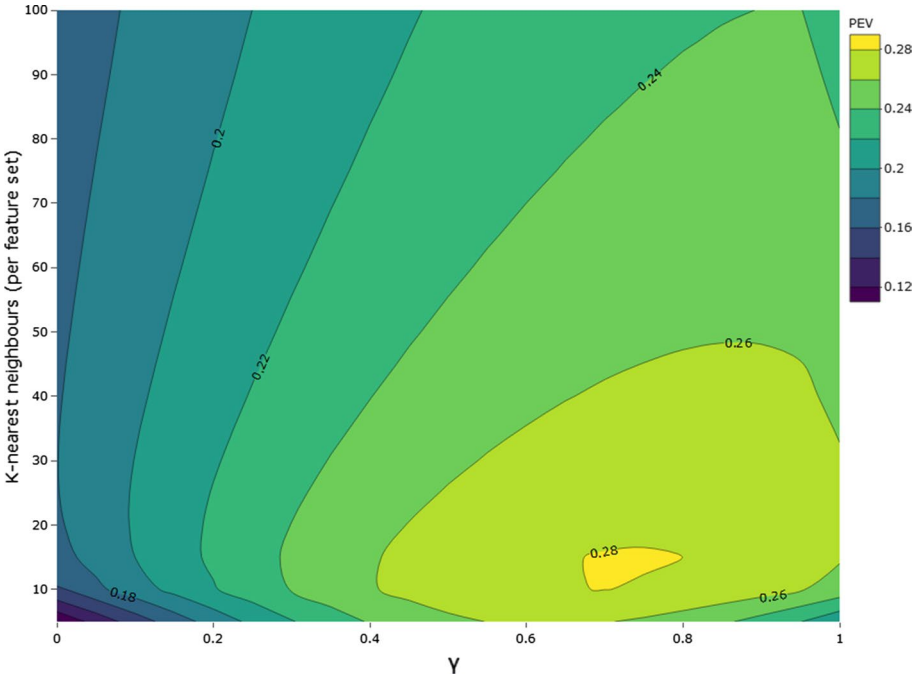

Fig. 2 Contour plot showing the influence of k-nearest neighbor and weighting parameter γ on PEV using weights from Eq. (6)

a classification scheme does not exist of course (if it would, we would indeed be using it for normalization). Instead, as a proxy, we consider here a subset of the articles for which arguable the most sophisticated item-level classification scheme is available, that is, articles from the medical domain that are classified with descriptors and subheadings from the Medical Subject Headings (MeSH) thesaurus. MeSH—a controlled vocabulary of biomedical descriptors that is used to describe the subject of articles in Medline–is created and updated by the US National Library of Medicine. The vocabulary contains more than 28 thousand MeSH descriptors that are arranged hierarchically by subject categories with more specific descriptors arranged beneath broader descriptors (National Library of Medicine 2019). MeSH descriptors can be designated as major indicating that they correspond to the major topics of the article, whereas non-major descriptors are added to reflect additional topics substantively discussed within the article. Further, approximately 80 subheadings (or qualifiers) can be used by the indexer to qualify a MeSH descriptor.

The effect of a particular citation normalization approach in this context can be assessed by looking at how the top $z\%$ of the articles are distributed over MeSH-based subject groups when the top $z\%$ is selected based on the normalized citation counts. Let $\hat{D} \subseteq D$ be the subset of articles with MeSH descriptors and $\left|\hat{D}\right| = \hat{N}$. Further, let these articles be partitioned into G different subject groups based on the MeSH classification (as explained below) and let $\hat{N}_g$ be the number of articles in the gth group. The size of the set of top $z\%$ articles is denoted by $\hat{n}_z$ (the number of articles with a normalized citation count above the $100-z$ percentile), which is normaly not excactly $z\%$ of all articles as there are ties in the normalized citation distribution. The expected number of articles from a given group if

only the size of the group matters–and therefore not the subject matter of these articles—is $\hat{n}_z/\hat{N} \times \hat{N}_g$. Deviations between the observed number of top z% articles in each subject group and the expected number of top z% articles can be considered as an indicator of how well the normalization approach corrects for subject matter when subject heterogeneous articles are compared.

In our case $\hat{N} = 456{,}045$, and we create to two cluster solutions of different granularity based on the subject similarity estimated from MeSH descriptors and subheadings. First, we calculate a weight (information content, IC) for each descriptor (Zhu et al. 2009). Let freq(desc$_i$) denote the frequency of descriptor i (here calculated over all articles in Medline with publication year 2009), then:

$$\mathrm{IC}(\mathrm{desc}_i) = -\log(P(\mathrm{desc}_i)) \tag{10}$$

where

$$P(\mathrm{desc}_i) = \frac{\mathrm{freq}(\mathrm{desc}_i) + \sum_{d \in \mathrm{descendants}(\mathrm{desc}_i)} \mathrm{freq}(d)}{\sum_{k=1}^{s} \left(\mathrm{freq}(\mathrm{desc}_k) + \sum_{d \in \mathrm{descendants}(\mathrm{desc}_k)} \mathrm{freq}(d) \right)} \tag{11}$$

We then represent each article by a vector of length $s + (s \times m)$ where s and m are the total number of unique MeSH descriptors and unique[3] number of subheadings in the dataset, respectivly. The vector position for the ith descriptor is given by $(m+1) \times i - m$ and the correspondig *weight for article* x_l ($\omega_i(x_l)$) is defined as

$$\omega_i(x_l) = \begin{cases} 0 & \text{if desc}_i \text{ is absent in } x_l \\ \mathrm{IC}(\mathrm{desc}_i) \times 1 & \text{if desc}_i \text{ is a minor descriptor in } x_l \\ \mathrm{IC}(\mathrm{desc}_i) \times 2 & \text{if desc}_i \text{ is a major descriptor in } x_l \end{cases} \tag{12}$$

The vector position for the jth subheading connected to the ith descriptor is given by $(m+1) \times i - m + j$ and the corresponding *weight for article* x_l ($\varphi_{ji}(x_l)$) is defined as

$$\varphi_{ji}(x_l) \begin{cases} 1 & \text{if subheading } j \text{ and descriptor } i \text{ are present in } x_l \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Note that many descriptor-subheading pairs are nonsensical and will never exist in practice and the subheading in such a pair will thus always take on the value 0 in the vectors.

Finally, then, the subject similarity between the articles in D is estimated by the cosine similarity between their corresponding vectors defined above. These similarity[4] estimates are then used as input to a clustering routine—modularity clustering, smart local moving algorithm (Waltman and van Eck 2012, 2013b)—to partition the articles into subject

[3] A group of MeSH descriptors that routinely are added to most articles, so called "check tags", are concepts of potential interest, regardless of the general subject content of the article (examples are "Human" and "Adult"). We do not include such check tags in any calculations.

[4] The top 25 most similar articles are identified for each article and the resulting non-symmetric matrix A is converted into a symmetric similarity matrix $B = A + A'$, where A' is the transpose of A. This is a requirement for the clustering routine.

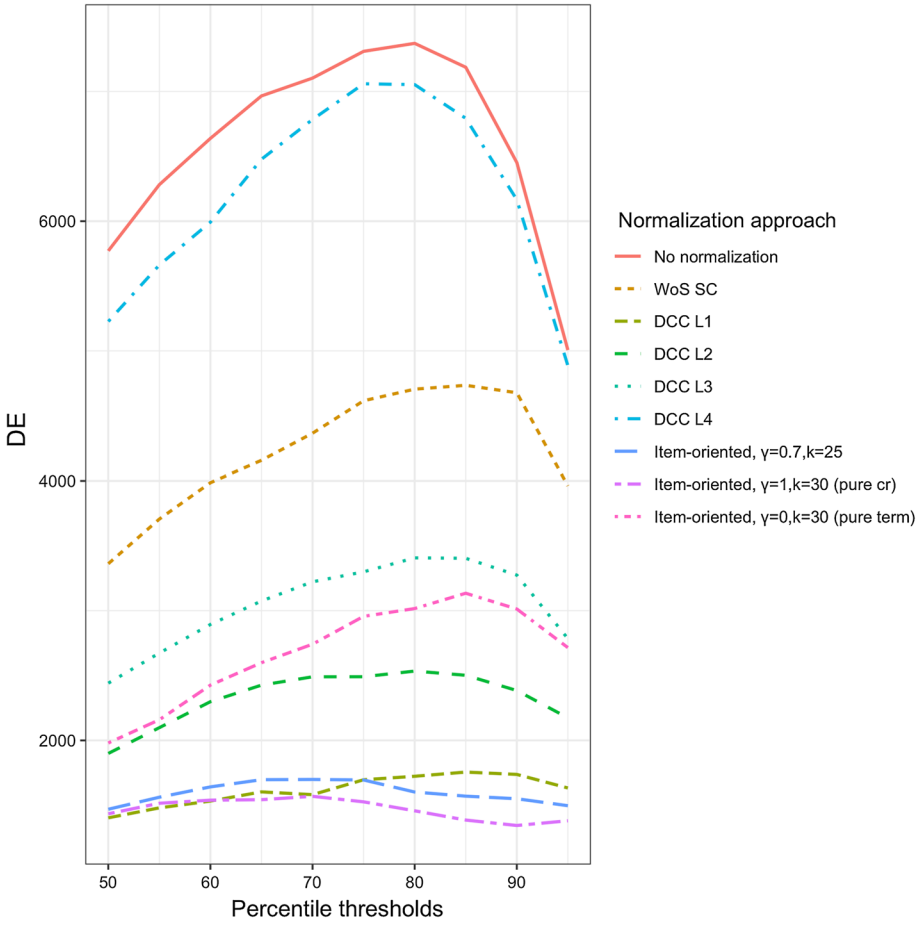

Fig. 3 Low resolution partition based on MeSH

coherent groupings. We create two partitions of different granularity: low resolution (resolution parameter equal to 2.0, 113 clusters) and high resolution (resolution parameter equal to 25.0, 1000 clusters).

We are now ready to give a precise description of the MeSH-based assessment of the normalization approaches. For a given partition (p), normalization approach (a) and percentile threshold (t), which defines the top $z\%$ articles, we summarize the deviations from expectations (DE) with

$$\mathrm{DE}(p,t,a) = \sum_{g \in p} \frac{(\mathrm{observed}_g^{p,t,a,} - \mathrm{expected}_g^{p,t,a})^2}{\mathrm{expected}_g^{p,t,a}} \tag{14}$$

where $\mathrm{observed}_g^{p,t,a,}$ is the number top $z\%$, i.e. top $(100-t)\%$, articles in cluster g under partition p when normalization approach a is used and $\mathrm{expected}_g^{p,t,a}$ the corresponding expectation. For the percentile threshold t, the following values were used: 0.50, 0.55, …, 0.95.

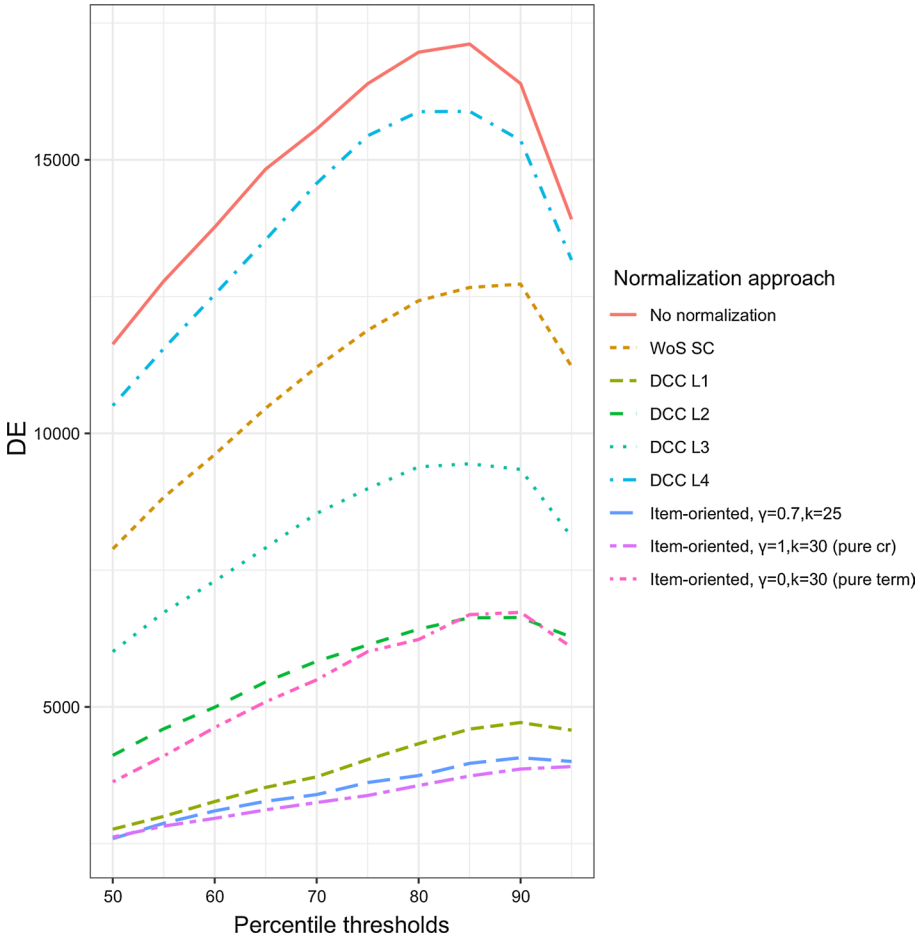

Fig. 4 High resolution partition based on MeSH

Figures 3 and 4 compare the effect the normalization approach has for different thresholds and for the low and high resolution partitions, respectively. Note that curves for non-normalization of citations are included in the two figures.

As can be seen, the patterns are similar for the two partitions (although the sum of deviations is much larger for the high resolution partition). For higher threshold values, the item-oriented approach with $\gamma = 1$ (i.e., a solution based only on bibliographic coupling) has slightly lower deviation values than the DCC L1 and the item-oriented solution based on hybrid similarity ($\gamma = 0.7$) in both partitions. That the item-oriented approach based only on cited references performs better than the item-oriented approach based on hybrid similarity is not completely surprising given Fig. 2, where it was shown that involving terms in the calculations in most cases decreased PEV values. However, it is not unreasonable to expect that heavily weighting of terms in the item-oriented approach should correlate with lower deviation values as terms from titles and abstracts are less independent from MeSH descriptors and subheadings than cited references. This is not the case, though. In fact, an item-oriented approach based only on terms ($\gamma = 0$) performs on pair with DCC

L2 normalization with respect to deviation under the high resolution partition. Under the low resolution partition, the pure term item-oriented approach has deviation values that put it between DCC L3 and DCC L2.

Considering DCC L4, one can see that it is a far too coarse system for normalization and its performance is just slightly better than no normalization at all. Normalization based on the approach WoS SC has a large effect compared to no normalization at all but is still far from the better performing approaches.

Only DCC L1 can compete with the item-oriented normalization. However, the item-oriented approach performs better (except the version that is based on terms only), especially at the higher thresholds.

Inadvertently effects and over-normalization

It is reasonable to worry if a normalization approach that aims to reduce the effect of subject matter in fact does more than that in such a way that it has unwanted and counterintuitive consequences. Especially the more sophisticated approaches that work on article-level and make use of citation linkage and/or textual content, might they indirectly adjust the raw citation counts with respect to other (non-subject matter) characteristics of the articles? While this deserves thorough investigation, here we will briefly shed some light on this question by taking advantage of the finer graded article type classification that are available in Medline compared to WoS. Using the subset of biomedical and life science articles, we investigate the ordering of aggregates of Medline article types based on the average number of received citation before any subject normalization is made. This ordering is as follows, from high average citations received to low average citations received:

Meta-Analysis > Randomized Controlled Trial > Clinical studies > Journal article (Other) > Comment/letter/editorial > Case Report.

This ordering based on average citations received is in line with what one would expect based on the characteristic of these publication types such as scientific rigor, i.e., randomized controlled trials versus non-randomized clinical studies or case reports and the degree of information content, i.e., meta-analysis versus single studies articles. When the ordering is based on citations that are normalized with respect to subject matter, the ordering should reasonably stay intact. E.g., it would be strange if, on average, case reports where now ranked higher than randomized controlled trials. As can be seen in Table 1, such a nonsensical effect of the considered normalization approaches is not observed. The rank order is intact, and the linear correlation is close (or equal) to 1.0 for all normalization approaches.

Discussion and conclusions

In this contribution, we have investigated two sophisticated article-level approaches to ex-post citation normalization. One of these approaches is item-oriented and has two parameters (k, γ), whereas the other falls under the algorithmically constructed classification system approach and has one parameter (the resolution parameter). For the latter approach, four sub approaches were investigated, where these differ only with regard to clustering solution granularity. We also included, as a benchmark, a traditional journal-based approach, WoS SC, in which the WoS journal subject categories were used. For the whole set of articles and for the assessment measure PEV (proportional reduction of

Table 1 Linear (rank-order, Kendall's tau used) correlation between average raw citations and average normalized citations for aggregates of Medline article types

WoS SC	DCC L1	DCC L2	DCC L3	DCC L4	Item-oriented $\gamma = 0.7, k = 25$	Item-oriented (pure cr, $k=30$)	Item-oriented (pure term, $k=30$)
0.998 (1.0)	0.979 (1.0)	0.994 (1.0)	0.997 (1.0)	1.000 (1.0)	0.976 (1.0)	0.972 (1.0)	0.992 (1.0)

variation in the citation distribution), the item-oriented approach had the best performance ($k \approx 25$, $\gamma \approx 0.7$), followed by the sub approach DCC L1, which corresponds to the most fine-grained clustering solution among the four solutions used in the study. DDC L4 and DDC L3, corresponding to low resolution clustering solutions, performed worse than or similar to WoS SC. For the MeSH-based assessment of the normalization approaches, in which the measure DE (deviation from expectation) was used, only DCC L1 could compete with the item-oriented approach. However, the item-oriented approach performed better (when cited references were heavily weighted in the similarity calculations), especially at the higher percentile thresholds.

Both evaluation methods we make use of to investigate the characteristics of the normalization approaches broadly paint the same picture with respect to the effectiveness of the approaches at correcting for subject matter effects on the amount of citations an article receives. The PEV method emphasizes that the reference set should contain highly subject similar publications with respect to the target publication whose citation count we seek to normalize and assumes that the degree to which this is the case correlates with the reference values ability to predict the observed citation count for the target publication. The DE method emphasizes that the citation distributions of different fields should differ as little as possible after normalization has been applied. Criterion such as these makes sense especially if citations are conceptualized as indicative of the influence a publication has (at a given snapshot in time) on surrounding research activities of comparable topicality, that is, in situations where we explicitly avoid differentiating between different specialties or scientific problem areas with respect to some notion of a hierarchy of importance. However, if citation counts are used within a theoretical framework that claims to assess scientific quality or importance in general, then normalization of the kind presented in this paper might be less appropriate.

A partial explanation for the observation that reference values given by the item-oriented approach that incorporate terms in the calculation of similarity estimates tend to perform worse than just using cited references might simply be that terms introduce more noise when publications from the same scientific problem area are to be identified compared with the case when cited references are used. While terms are connected to the communicative aspect of fields and specialties as it captures specific terminology, cited references connects to the cognitive aspects of a given scientific problem area as they mirror a shared body of theories, methods and important papers (Rons 2018). Subject matter mismatch is probably more likely to happen when terms (as are far less specific than cited references) are used compared to when cited references are used for similarity estimation, especially using a quite simple and straightforward bag-of-terms approach as we do here.

The item-oriented approach generally performs better according to PEV and DE than the DCC approaches and this can be attributed to the supposition that the item-oriented approach probably is less likely to create reference sets that mostly include the output of a few authors citing each other–isolated from the scientific communication that goes on in their field of study–and such small clusters might be unable to form reasonable reference values (Waltman and van Eck 2013a). This has more to do, though, with the use of direct citations than with the clustering approach as such. That the DCC approaches by design ignore the interrelatedness and overlap of fields and specialties while the item-oriented approach does not, is also likely to play a role in observed differences in their effectiveness. Small reference sets, as both the item-oriented approach and the DCC approaches can create are often questioned, since it can lead to situations where a publication is mainly compared with itself. However, this is only true when the target publication is itself a member of its own reference group. Note that we do not allow a publications' citation count

to influence its own reference value. Our approach is non-standard in this regard but the praxis of including the target publication in its own reference group probably stems from the fact that traditional journal-based reference groups generally are very big and, in such cases, it hardly matters if we include the publication or not (but including it simplifies calculations). For the publication-level approaches we consider, though, this issue is very important as including the target publication in its own reference set could be problematic.

With respect to small or highly specific reference sets one can note that in a different but similar context of identifying relatedness measures for accurate clustering solutions of scientific articles, it has been shown that rather small values of k in top-k most similar publications yield more accurate clustering solutions than higher values of k (Waltman et al. 2019). This finding can be said to be mirrored in the item-oriented approach and indicted in the DCC approaches.

We also note that, at least tentatively and on an aggregated level, that these publication-level and fine-grained approaches do not seem introduce any adverse effects as they preserve the common-sense citation ranking for publication types intact (e.g., randomized trials versus case studies).

While both publication-level approaches echoes the reasoning of Kostoff and Martinez (2005) that the only meaningful normalization approach is to select for each publication a small number of thematically similar publications and to compare the number of citations of a publication with the number of citations received by the selected similar publications, the item-oriented approach is arguably closer to this view than the DCC approaches. The DCC approaches use of direct citations is partly a result of efficiency restrictions with respect to implementation aspects. In principle, however, nothing hinders that direct citation is replaced or complemented with other data for similarity calculations and this would potentially increase its normalization effectiveness. One of the main strengths of the item-oriented approach is that it avoids to artificially place each publication in exactly one specialty, field or whatever notion of organizational and topical entities one tries to define and uses as a basis for the reference set construction. However, this strength might sometime become a weakness as it can be harder to communicate the result when one cannot make references to a classification system with labeled groups (even if it is possible to automatically generate content labels from the publications in a created reference set, as indeed has been done with respect to the four-level hierarchical classification system used in this study). The aim and context of the exercise in which normalized citation values are used will determine when the advantages outweigh the disadvantages.

As we have restricted the investigation to the ex-post family of normalization approaches, further inquiry is required to investigate how the best performing approaches we have studied fair against the best ex-ante approaches. Though this is an empirical question, there are some evidence suggesting that ex-ante approaches would not perform better than the item-oriented approach or a fine-graded DCC solution. Earlier research mostly identified ex-post approaches as being more efficient at reducing the variability in received citations due to subject matter heterogeneity. However, these studies have been criticized by Sirtes (2012) as being biased in favor of ex-post approaches and hence one would be advised not to put too much emphasis on these findings. Ex-ante approaches seem to have a small advantage over a particular instance of ex-post normalization when the bias in the evaluation methodology is removed (Waltman and van Eck 2013a, b) though this is somewhat disputed by Ruiz-Castillo (2014). What is important here is that when ex-ante approaches have been shown to perform slightly better than ex-post approaches, the ex-post normalization considered have been based on WoS journal subject categories. As we have seen in this study, the item-oriented approach and the fine-graded DCC solutions by far

surpass normalization based on these journal subject categories, making it unlikely that the best performing ex-ante approach would come out on top. This is an issue for future studies, however.

Finally, in future studies it might be of interest to investigate the correlation between ratings derived by these normalization approaches and peer rating. However, it is not obvious that peer review ratings are the ground truth against which these approaches should be assessed. The two methods might have quite different goals, especially if the conceptualization of the meaning of citations is a more modest one that does not extend beyond scientific peer influence at a given time. The fact that the reliability of peer review is not necessarily high and that the chance factor in peer review outcomes can be quite substantial also make such an investigation challenging, but nevertheless interesting.

References

Adams, J., Gurney, K., & Jackson, L. (2008). Calibrating the zoom—A test of Zitt's hypothesis. *Scientometrics, 75*(1), 81–95.

Braun, T., & Glänzel, W. (1990). United Germany—The new scientific superpower. *Scientometrics, 19*(5–6), 513–521.

Colliander, C. (2015). A novel approach to citation normalization: A similarity-based method for creating reference sets. *Journal of the Association for Information Science and Technology, 66*(3), 489–500.

Colliander, C., & Ahlgren, P. (2012). Experimental comparison of first and second-order similarities in a scientometric context. *Scientometrics, 90,* 675–685.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLoS ONE, 8*(3), e58727.

CWTS Leiden Ranking 2015 Methodology. (2015). Retrieved from http://www.leidenranking.com/Content/CWTS%20Leiden%20Ranking%202015.pdf. Accessed 13 May 2019.

Glänzel, W., & Moed, F. H. (2013). Opinion paper: thoughts and facts on bibliometric indicators. *Scientometrics, 96*(1), 381–394.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori versus a posteriori normalisation of citation indicators. The case of journal ranking. *Scientometrics, 87*(2), 415–424.

Kostoff, R. N., & Martinez, W. L. (2005). Is citation normalization realistic? *Journal of Information Science, 31*(1), 57–61.

Leydesdorff, L., & Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology, 62*(2), 217–229.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., & De Nooy, W. (2013). Field-normalized impact factors (IFs): A comparison of rescaling and fractionally counted IFs. *Journal of the American Society for Information Science and Technology, 64*(11), 2299–2309.

Li, Y., Castellano, C., Radicchi, F., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics, 7*(3), 746–755.

Li, Y., & Ruiz-Castillo, J. (2013). The comparison of normalization procedures based on different classification systems. *Journal of Informetrics, 7*(4), 945–958.

Moed, H. F., De Bruin, R. E., & van Leeuwen, T. N. (1995). New bibliometric tools for the assessment ofnational research performance: Database description, overview of indicators and first applications. *Scientometrics, 33*(3), 381–422.

Neuhaus, C., & Daniel, H. D. (2009). A new reference standard for citation analysis in chemistry and related fields based on the sections of chemical abstracts. *Scientometrics, 78*(2), 219–229.

Newman, M. E. J. (2004a). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*(6), 066133.

Newman, M. E. J. (2004b). Analysis of weighted networks. *Physical Review E, 70*(5), 056131.
Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2017). A comparison of the Web of Science with publication-level classification systems of science. *Journal of Informetrics, 11*(1), 32–45.
Radicchi, F., & Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics, 6*(1), 121–130.
Rons, N. (2018). Bibliometric approximation of a scientific specialty by combining key sources, title words, authors and references. *Journal of Informetrics, 12*(1), 113–132.
Ruiz-Castillo, J. (2014). The comparison of classification-system-based normalization procedures with source normalization alternatives in Waltman and Van Eck (2013). *Journal of Informetrics, 8*(1), 25–28.
Sirtes, D. (2012). Finding the easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics, 6*(3), 448–450.
Thelwall, M. (2019). The influence of highly cited papers on field normalised indicators. *Scientometrics, 118*(2), 519–537.
U.S. National Library of Medicine. (2019). Principles of MEDLINE Subject Indexing. Retrieved from https://www.nlm.nih.gov/mesh/introduction.html. Accessed 13 May 2019.
van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE, 8*(4), e62395.
van Leeuwen, T. N., & Medina, C. C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics. *Research Evaluation, 21*(1), 61–70.
van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*(3), 397–420.
Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2019). A principled methodology for comparing relatedness measures for clustering publications. *arXiv e-prints*. Retrieved from https://arxiv.org/abs/1901.06815.
Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378–2392.
Waltman, L., & van Eck, N. J. (2013a). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics, 7*(4), 833–849.
Waltman, L., & van Eck, N. J. (2013b). A smart local moving algorithm for large-scale modularity-based community detection. *The European physical journal B, 86*(11), 471.
Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011a). Towards a new crown indicator: an empirical analysis. *Scientometrics, 87*(3), 467–481.
Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011b). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37–47.
Zhu, S., Zeng, J., & Mamitsuka, H. (2009). Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. *Bioinformatics, 25*(15), 1944–1951.
Zitt, M. (2010). Citing-side normalization of journal impact: A robust variant of the audience factor. *Journal of Informetrics, 4*(3), 392–406.
Zitt, M. (2013). Variability of citation behavior between scientific fields and the normalization problem: The "citing-side" normalization in context. *Collnet Journal of Scientometrics and Information Management, 7*(1), 55–67.
Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics, 63*(2), 373–401.
Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology, 59*(11), 1856–1860.