Transforming the Heterogeneity of Subject Categories into a Stability Interval of the MNCS

Marion Schmidt[1] and Daniel Sirtes[1*]

[1] *schmidt@forschungsinfo.de, sirtes@forschungsinfo.de*
iFQ Institute for Research Information and Quality Assurance, Schützenstr. 6a, D-10117 Berlin (Germany)

Abstract

The internal homogeneity of research disciplines in subject categories (SC) of the Web of Science database (WoS) regarding their publication and citation practices is an essential precondition for the field-normalization of citation indicators. This imperative of underlying homogeneity seems not to be met throughout all categories, as has been shown in former research. A keyword-based clustering method displays both the diversity of research areas included in an SC and that the clusters' mean citation rate differ substantially. This proof-of-concept paper on the basis of one country set and two SCs presents a bootstrapping method, which allows quantifying the degree of heterogeneity within subject categories as a stability interval. The MNCS 95% stability interval of our set has a range of 6.7% and 7.3% compared to its score. This kind of robustness measure could be implemented for future evaluative citation analysis in order to convey the coarseness of bibliometric point estimates.

Conference Topics

Methods and techniques; Citation and co-citation analysis; Indicators

Introduction

Field-normalized citation indicators such as the MNCS (Waltman, Eck, Leeuwen, Visser, & Raan, 2011) normalize the citation rate of a given publication corpus based on expectancy values of subject categories which correspond to the respective average citation rates within a research field (Vinkler, 1986; Mcallister, Narin, & Corrigan, 1983). Field normalization has been developed in order to neutralize the obvious diversity of publication and citation practices between field and subfields, as a corrective to otherwise unfair comparisons between the citation impact results of corpora with varying subject distributions.

Various methods for field delineation have been proposed (Glänzel & Schubert, 2003; Glänzel, Thijs, Schubert, & Debackere, 2009; Ruiz-Castillo & Waltman, 2014) including many proposals for clustering methods and arguments to determine the correct levels of aggregation. So far, however, no classification systems other than those provided by the database vendors could be established as standard throughout the bibliometrics community.

However, it is easily observable that the classification of the WoS subject categories diverges in size and specificity. Van Eck et al. (2013) provide furthermore strong evidence of heterogeneity within the medical subject categories along the characteristics of clinical and experimental research: After terms have been extracted from titles and abstracts, substructures are made visible by a term cloud procedure. These substructures can be assigned intellectually to clinical or experimental research and differ significantly in their citation rates along these dimensions. An intuitive explanation for this phenomenon would be the assumption that clinical researchers cite experimental studies, but that experimental researchers cite clinical studies only to a lesser extent.

Van Eck et al. (2013) draw the conclusion that the impact of clinical research is structurally underestimated by classical normalized citation indicators. The substructures made visible correspond to a facet that can be seen as transverse to a valid and comprehensible classification according to medical fields such as Clinical Neurology, Cardiac &

[*] The order of authorship is merely alphabetical.

Cardiovascular Fields, etc. Further theoretical issues beyond classification or clustering criteria seem to be not yet solved: If, for example, publications in so called hot topic areas are compared only with similar publications, even only with those who share not only the same topic, but also the same instruments, etc.? This could be seen as an over-normalization (Sirtes, 2012b; Sirtes, 2012a). Or is it legitimate to aggregate hot topics with less active research areas and thereby highlight the former as particularly successful? With the latter attitude the strategic decision of a researcher for a high impact research fields would be gratified while at the same time an implicit premise would be set that not all delineable areas in a functionally differentiated research landscape would be of equal value, insofar impact differences, which are effects of the functional differentiation, would not be neutralized.

By introducing finer classification systems these issues are addressed, although not answered based on theoretical reasons, as only further normalization options are created, whereas the resulting differences are not directly interpretable. Besides, in-house classifications systems are not easily compatible with a desirable trend towards greater standardization and reproducibility in the bibliometric community.

In the present paper we introduce a concept for quantifying heterogeneity differences within subject categories and thus maintain the WoS subject categories as basis for the field normalization, as they provide community-wide comparability and mutual reproducibility. Heterogeneity differences between subject categories are quantified and used to construct error or stability intervals, which can be integrated into the calculation of the total impacts of an institution or a country as before. The approach thus combines two advantages: on the one hand, we continue to work at the level of a standard classification system and on the other hand, underlying structures on a secondary level are made transparent.

Methods and Data

Keyword terms of all articles, reviews and letters published in journals of two medical subject categories (Parasitology (P), Otorhinolaryngology (O)) of the publication year 2008 have been extracted.[1] WoS keywords are not a controlled vocabulary like, e.g., Medical Subject Headings in PubMed/Medline and are therefore not per se complete and normalized. Table 1, however, shows that the amount of publications that have not assessed with keywords is relatively small. Keywords have, on the other hand, the advantage of simple accessibility; it is not necessary to exclude i.e. filler words. In order to accomplish a basic normalization, a stemming procedure is carried out which neutralizes different inflexions.

All distinct keyword terms are normalized with an Oracle Text stemming function and coupled by the *contains* function, again as provided in Oracle Text. Stemmed terms must therefore not be necessarily identical, but one term can contain the other, respectively. This also applies to keywords, which are phrases and may contain single keywords and be thus coupled with them. These keyword pairs are used for a coupling procedure of the corresponding publications; Salton’s Cosine is used to neutralize differing amounts of keywords.

With the aim to reproduce the visual substructures of Van Eck et al. (2013) in a first step with our cluster procedure, these two subject categories have been chosen as they display different types of sub-structures in the discussed work. Parasitology displays quite distinct structures with three visible clusters seemingly characterized by significant differences in citation levels whereas Otorhinolaryngology displays a more fuzzy structure.[2]

[1] All calculations are processed in an Oracle database of WoS raw data (SCI, SSCI, A&HCI, CPCI-S, CPCI-SS) frozen in the 17th calender week 2013.

[2] http://www.neesjanvaneck.nl/basic_vs_clinical/

Table 1: Share of publications with keywords.

	Parasitology	*Otorhinolaryngology*
JARL 2008 (all)	3727	5122
JARL 2008 (percentage of publications with keywords)	98.0%	90.6%

The ratio of realized to theoretical possible relations between all items gives an impression about the broadness of the empirical basis of the coupling results. Table Table 2 gives the percentage of realized to theoretically possible relations of all publications (JARL = Articles, Letters and Reviews with publication type Journal Article) in 2008.

Table 2: Ratio realized relations to possible relations.

	Parasitology	*Otorhinolaryngology*
JARL 2008 (all)	18.2%	11.3%
JARL 2008 (only with keywords)	19.0%	13.8%

The resulting distance measures for publication pairs are imported into the statistical program R, converted into dissimilarity values and the clustering method Ward is used. Ward as a standard hierarchical-agglomerative clustering procedure was chosen, because it is crucial for our approach to have a clustering procedure which does not require a fixed number of clusters as parameter. Furthermore, single linkage with its well-known tendency to dilated cluster structures seems to impose to weak requirements on the clusters' homogeneity and complete linkage too strong requirements.

The usual cut-off-value of 5 was determined manually; however in future iterations of the procedure the optimal cut off value will be estimated.

As shown in Table 2 not all publications in the respective sets are actually assigned with keywords, thus we have added a non-keyword cluster with its mean citation rate in order to represent all publications in our dataset. This appears as a legitimate solution given that fact that non-keyword items have considerably smaller mean citation rates compared to the whole subject category and have to be taken into account in order to appropriately represent the SC.

Results

The visualization for the subject category parasitology as resulting from (Van Eck et al.., 2013) indicates a distribution of three discernable substructures which are clearly different in citation level. With our method, we arrive at eleven clusters. Table 3 shows four of the top keywords[3] and the respective mean citation rates, whereas Figure 1 gives the frequency distribution of the clusters (as the width of the bars) and the mean citation rates in a histogram. The topics of the clusters can only partially confirm Van Eck et al.'s conclusion. The keywords of cluster 5, 6, and 7 have all clear connection to experimental laboratory research, however only 5 (with the most distinctly molecular biology focus) has a very high citation rate compared to the rest. It is possible, that parasitology is rather a special case compared to other medical SCs, as it also encompasses topics such as classical biology (cluster 1), epidemiology (clusters 2 and the more clinical 4), a veterinary cluster (8), and clusters that are joined by common parasites (3, 9,10, and 11).

[3] All keywords were in the top 10 most frequent ones. Redundant keywords (like 'plasmodium' and 'plasmodium falciparum') and keywords that were not informative in understanding the topic of the cluster (like 'parasites') were excluded.

Table 3 - Top keywords and mean citation rate of keyword clusters in parasitology (ordered by cluster size).

Cluster	Top Keywords				Mean Citation Rate
1	Phylogeny	Evolution	Ecology	Morphology	3.91
2	Infection	epidemiology	Seroprevalence	Antibodies	5.76
3	Malaria	plasmodium falciparum	infected erythrocytes	cerebral malaria	6.25
4	Transmission	Children	Resistance	Efficacy	7.02
5	Expression	in-vitro	Protein	gene-expression	7.57
6	Mice	in-vivo	dendritic cells	immune-response	6.69
7	Identification	PCR	linked-immunosorbent-assay	Antibodies	5.50
8	Sheep	Cattle	haemonchus-contortus	Ivermectin	4.11
9	Disease	trypanosoma cruzi	chagas disease	risk-factors	6.09
10	Diptera	Culicidae	aedes-aegypti	anopheles-gambiae	5.32
11	Cryptosporidium	Parvum	Giardia	Genotypes	7.88

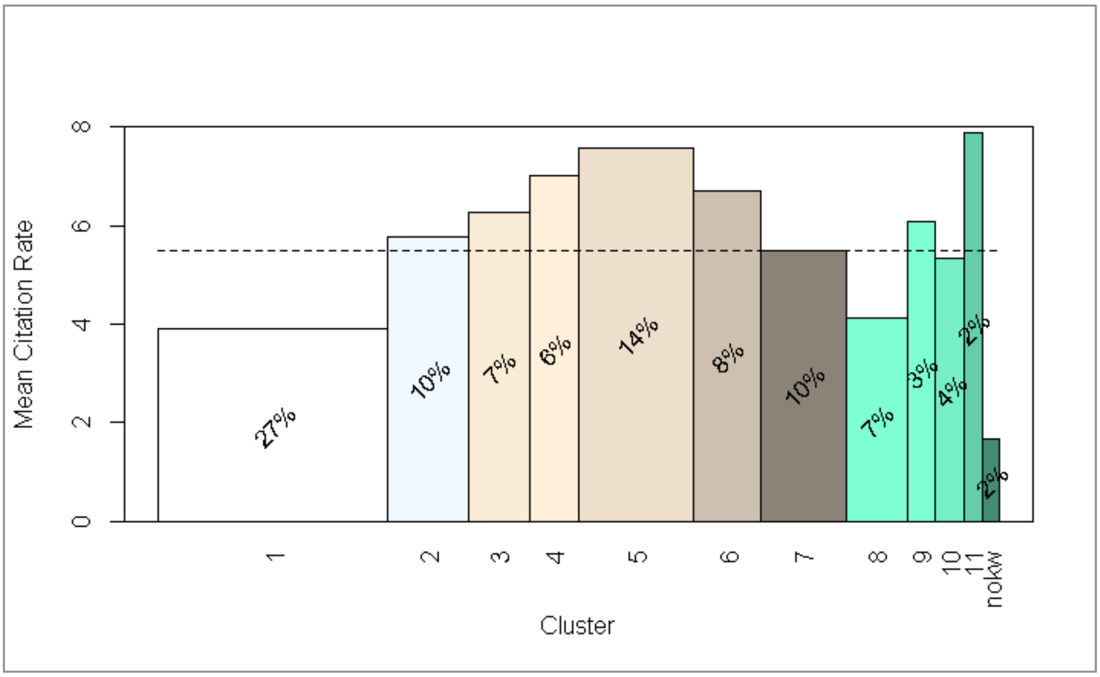

Figure 1: Share and Mean Citation Rate of Parasitology Clusters. The dotted line represents the MCR of the whole SC.

In the second case of otorhinolaryngology, the structure shown by (Van Eck u. a., 2013) is quite fuzzy and less-structured, which is mirrored by our cluster distribution. It consists of one larger and a considerable amount of very small cluster. There are also significant variations between mean citation levels ranging from around 2 to larger than 4, it is however more difficult to interpret the cluster's respective keyword frequencies.

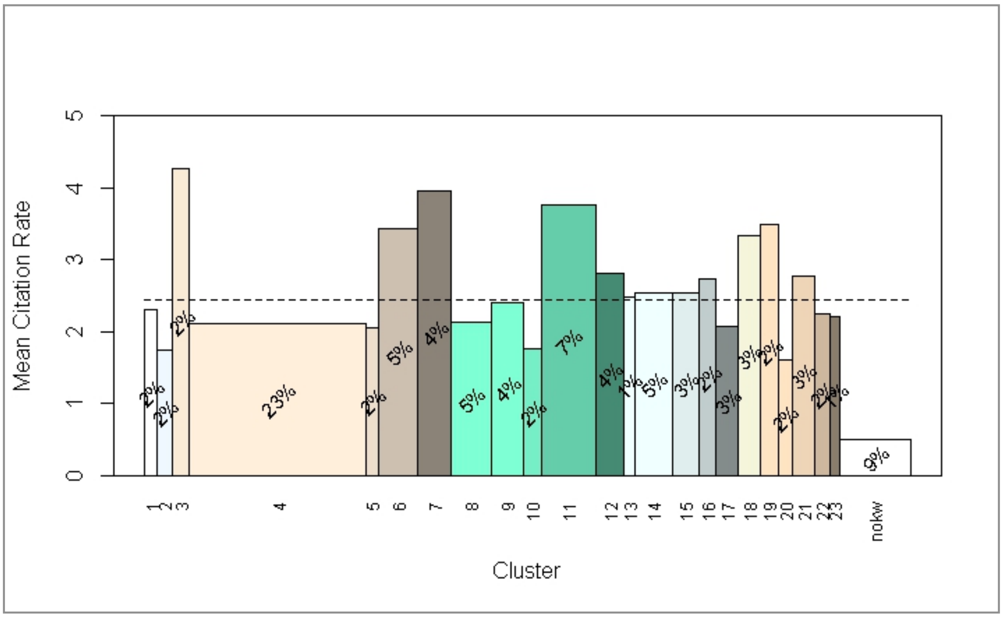

Figure 2: Share and Mean Citation Rate of Otorhinolaryngology Clusters. The dotted line represents the MCR of the whole SC.

In order to calculate the MNCS and its stability, sets of publications with an affiliation in Germany have been selected. The size of the sets were 208 (P) and 486 (O) publications respectively.

On the basis of the resulting cluster distributions, a bootstrapping approach has been utilized.

A set of MCR clusters equal to the size of the German set has been drawn with replacement from the clusters' MCRs with the probabilities equal to the clusters' share. The arithmetic mean of this combination has been calculated and served as the Expected Citation Score (ECS_i). Each raw citation score of the German papers was then divided by the ECS_i and the arithmetic mean of the results delivered the $MNCS_i$. 10'000 iterations of this procedure have been executed. The distribution of the scores are depicted in Figure 3.

Finally, the 2.5% and 97.5% quantiles of this distribution have been calculated.

The resulting MNCS 95% stability interval of the German set for parasitology ranges from 1.35 to 1.46 with an MNCS of 1.40 and for otorhinolaryngology from 0.87 to 0.93 with an MNCS of 0.9. Thus, although parasitology displays a much wider distribution, as can also be seen in Figure 3, the relative deviance of the MNCS ([95% range of $MNCS_i$]/MNCS) is quite similar with 7.3% and 6.7%, respectively.

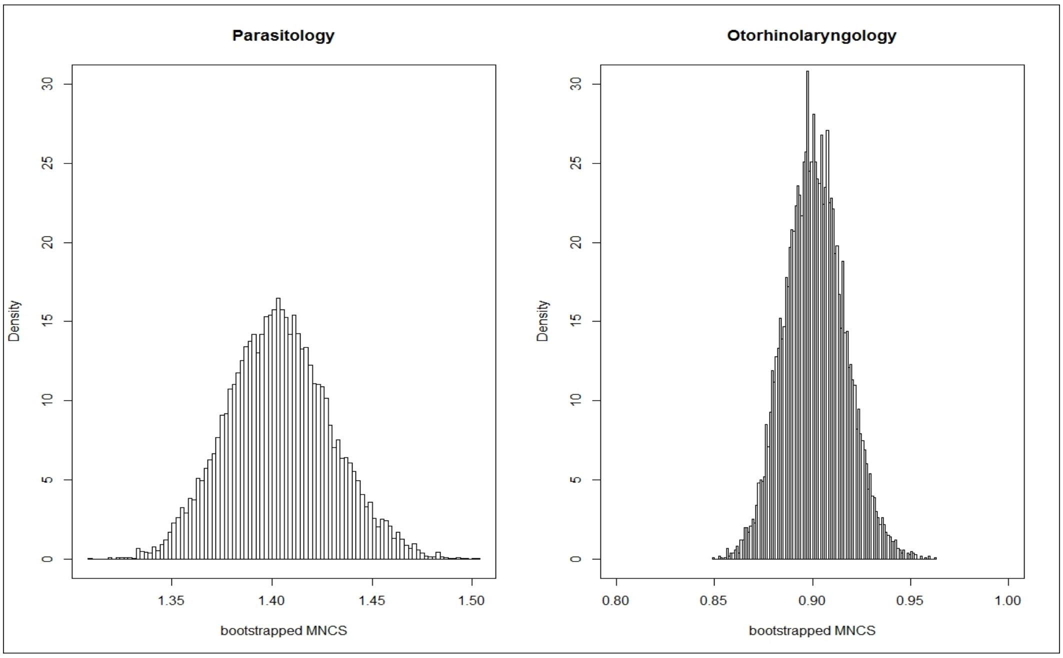

Figure 3: Distribution of MNCS$_i$ for German publications.

Discussion

These preliminary results show in the case of parasitology that clusters can be delineated and differing topical foci can be identified as well. While a dimension clinical versus experimental research is perceivable, other facets also occur: It may be the case that parasitology is a special SC as the clusters have also rather unusual topics compared to other medical disciplines such as classical biology, veterinary sciences and epidemiology. The Mean Citation Rates vary massively with a total range of MCRs of 3.97 citations per publication In the second case of otorhinolaryngology, the cluster distribution is less harmonic, more frayed out and not easily interpretable (confirming here the results of (Van Eck et al., 2013)). The coupling procedure succeeded on a relatively smaller amount of publications and many more clusters have been created. Furthermore, the citation levels are all much lower and the range of MCRs, the publications without keywords notwithstanding, have only a total range of 2.6 citations per publication.

The hitherto work was intended as a proof of concept: We were able to show that subject category substructures with different citation levels exist. Differences in citation homogeneity are however not in both cases concordantly attributable to topical structures. For the current state of this work, some simplifications have been applied: Citation rates should be processed and normalized document type-specific as articles, letters and reviews are cited differently. However, citation level differences in our results are so clear and dominant that they couldn't possibly only be caused by different document type patterns in the clusters. For a final implementation of this method, the calculations will be processed document type-specific and the expansion of the method to sets of multiple SCs, including an SC fractionalization will be developed. An exclusion of letters might be contemplated as for example about half of the publications without keywords in otorhinolaryngology are letters (about three quarters of all letters in this SC). Furthermore, parameters of the study like the clustering method and definition of cut off-values will be systematically varied and analyzed. It is even conceivable to calculate such stability intervals on the basis of percentile based indicators, which are less sensitive to outliers than the MNCS. However, already as it stands this method shows promise in circumventing to problem of calculating normalized citation scores on non-standard classification schemes while taking into account the heterogeneity of research areas in the classical WoS SC classification. This method could also be combined with already existing bootstrapping methods of the publications sets themselves as implemented for example in the Leiden Ranking (www.leidenranking.com). Together they could account for both the robustness of the citation scores given the size and distribution of the publication sets themselves, as well as the underlying uncertainty of the expected citation rates. We believe that such methods that display the coarseness of bibliometric point estimates, which especially clients of evaluative bibliometric analyses are prone to disregard and thus revel or despair at minute changes of their scores and ranks, are an important step to the correct interpretation of bibliometric indicators and crucial for the development of bibliometrics into a mature science.

References

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367. http://doi.org/10.1023/A:1022378804087

Glänzel, W., Thijs, B., Schubert, A., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, *78*(1), 165–188. http://doi.org/10.1007/s11192-008-2109-5

Mcallister, P. R., Narin, F., & Corrigan, J. G. (1983). Programmatic evaluation and comparison based on standardized citation scores. *IEEE Transactions on Engineering Management EM*, *30*(4), 205–211. http://doi.org/10.1109/TEM.1983.6448622

Ruiz-Castillo, J., & Waltman, L. (2014). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Working Paper*. Abgerufen von http://e-archivo.uc3m.es/handle/10016/18385

Sirtes, D. (2012a). 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.

Sirtes, D. (2012b). How (dis-)similar are different citation normalizations and the fractional citation indicator? (And How it can be Improved). In *Proceedings of 17th International Conference on Science and Technology Indicators* (S. 894–896). Montréal: Éric Archambault, Yves Gingras, and Vincent Larivière.

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. http://doi.org/10.1371/journal.pone.0062395

Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, *10*(3-4), 157–177. http://doi.org/10.1007/BF02026039

Waltman, L., Eck, N. J. van, Leeuwen, T. N. van, Visser, M.S., & Raan, A. F. J. van. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics*, *87*(3), 467–481. http://doi.org/10.1007/s11192-011-0354-5