# Science mapping and research evaluation

## A novel methodology for creating normalized citation indicators and estimating their stability

**Cristian Colliander**

**Department of Sociology**
PhD Thesis 2014

This work is protected by the Swedish Copyright Legislation (Act 1960:729)
ISBN: 978-91-7601-134-8
ISSN: 1104-2508
Electronic version available at http://umu.diva-portal.org/
Printed by: Print & Media, Umeå, 2014

*To my family*

# Table of Contents


**Table of Contents** **i**
**List of original articles in the thesis** **ii**
**Acknowledgements** **iii**
**Abstract** **iv**
**Introduction** **1**
Theoretical and empirical framing of citations 2
Pressing issues in the construction of citation indicators 6
**Background for the articles and general problem statements** **8**
Bibliometric identification of subject-related documents 9
Normalizing raw citation impact with respect to subject matter 12
Uncertainty and robustness of citation indicators 15
**Aim of the thesis** **18**
**Results: Summary of the four articles** **19**
ARTICLE I: Document–document similarity approaches and science mapping: Experimental comparison of five approaches 19
Article II: Experimental comparison of first and second-order similarities in a scientometric context 20
ARTICLE III: A novel approach to citation normalization: a similarity-based method for creating reference sets 20
ARTICLE IV: The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments 21
**Concluding discussion** **22**
**References** **26**

# List of original articles in the thesis

I. Ahlgren, P., & Colliander, C. (2009). Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics, 3*(1), 49-63.

II. Colliander, C., & Ahlgren, P. (2012). Experimental comparison of first and second-order similarities in a scientometric context. *Scientometrics, 90*(2), 675-685.

III. Colliander, C. (2014). A novel approach to citation normalization: A similarity-based method for creating reference sets. *Journal of the Association for Information Science and Technology*. Advance online publication. doi: 10.1002/asi.23193

IV. Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics, 5*(1), 101-113.

# Acknowledgements

Supervisors: Rickard Danell and Olle Persson

Partner in crime: Per Ahlgren

Sorcerer: Simon Lindgren

Shaman: Ragnar Lundström

# Abstract

The purpose of this thesis is to contribute to the methodology at the intersection of relational and evaluative bibliometrics. Experimental investigations are presented that address the question of how we can most successfully produce estimates of the subject similarity between documents. The results from these investigations are then explored in the context of citation-based research evaluations in an effort to enhance existing citation normalization methods that are used to enable comparisons of subject-disparate documents with respect to their relative impact or perceived utility. This thesis also suggests and explores an approach for revealing the uncertainty and stability (or lack thereof) coupled with different kinds of citation indicators. This suggestion is motivated by the specific nature of the bibliographic data and the data collection process utilized in citation-based evaluation studies.

The results of these investigations suggest that similarity-detection methods that take a global view of the problem of identifying similar documents are more successful in solving the problem than conventional methods that are more local in scope. These results are important for all applications that require subject similarity estimates between documents. Here these insights are specifically adopted in an effort to create a novel citation normalization approach that – compared to current best practice – is more in tune with the idea of controlling for subject matter when thematically different documents are assessed with respect to impact or perceived utility. The normalization approach is flexible with respect to the size of the normalization baseline and enables a fuzzy partition of the scientific literature. It is shown that this approach is more successful than currently applied normalization approaches in reducing the variability in the observed citation distribution that stems from the variability in the articles' addressed subject matter. In addition, the suggested approach can enhance the interpretability of normalized citation counts. Finally, the proposed method for assessing the stability of citation indicators stresses that small alterations that could be artifacts from the data collection and preparation steps can have a significant influence on the picture that is painted by the citation indicator. Therefore, providing stability intervals around derived indicators prevents unfounded conclusions that otherwise could have unwanted policy implications.

Together, the new normalization approach and the method for assessing the stability of citation indicators have the potential to enable fairer bibliometric evaluative exercises and more cautious interpretations of citation indicators.

# Introduction

Performance-based university research funding systems that incorporate bibliometric measures are operational in several countries (Hicks, 2012). Alongside these national evaluation systems, metrics derived from publication data are increasingly used at the level of institutions or departments for performance reviews, tenure decisions, and similar purposes (Abbott et al., 2010). It can now be considered standard practice for an evaluation report of an institution to include the number of publications and the number of citations these publications have received, at least in the natural and life sciences (Bornmann, 2013). The increasing uses of bibliometric exercises that have real consequences for the entities subjected to them also increases the importance that such exercises are valid and that the outcomes of bibliometric investigations are not over-interpreted.

Bibliometrics offers a large set of quantitative methods and measures for studying the structure and process of formal scholarly and scientific communication. Because this communication is realized through publications, scientific and scholarly explanations and knowledge claims – along with their reception, diffusion, and interrelations – can be illuminated by examining the documents that represent the important outcomes from different research endeavors (Morris & Van der Veer Martens, 2008). It is necessary, therefore, that a research publication is embedded within a community-generated body of literature for its potential relevance and importance to be demonstrable.

A distinction is usually made between relational bibliometrics and evaluative bibliometrics (Borgman & Furner, 2002). In the former case, indicators of the strength of the relationship, or the direction of flow, between documents, authors, journals, research communities, organizations, or nations are in focus. The main aim of relational bibliometrics is to map social aspects and/or the manifestations of cognitive productions in different scientific problem areas (Borner, Chen, & Boyack, 2003; White & McCain, 1997) or to assist in information retrieval tasks (Wolfram, 2003). The use of bibliometrics for evaluation focuses on deriving impact indicators for different units of assessment such as individual researchers, departments, journals, or aggregates thereof and examining the influence these entities have upon the associated research activity. The main aim of evaluative bibliometrics is to assess different aspects of research performance (Moed, 2005; Narin, 1976).

This thesis is situated at the intersection of relational and evaluative bibliometrics. The contribution of this thesis revolves around issues of how to derive estimates of the subject similarity between documents and how we can use such information to create frames of references in which raw citation counts can be contextualized. This will enable investigations into the degree to which different scientific documents influence their respective fields of inquiry. It follows, therefore, that the evaluative framework will be characterized by the analysis of citations. The reasoning and discussion herein are thus limited to research areas that can be characterized by the standard mode of formal communication in the natural sciences, in other words, those research areas where articles in international journals are the main form of communication. This is primarily because modern data sources for citation analysis, i.e., comprehensive bibliographic citation indices such as those provided by Thomson Reuters, do not adequately cover the research output from potential units of assessment where other communications channels are of more importance. Estimates of the significance of journal publications and the coverage of this literature in standard citation indices for different areas of science and scholarship are given, for example, by Moed (2005, ch. 7) and by Sivertsen and Larsen (2012).

## Theoretical and empirical framing of citations

Notwithstanding literature coverage issues, evaluative citation analysis is coupled with controversies of a more conceptual nature. Several different theoretical interpretations of the "meaning" of citations are available. The so-called "normative view" builds on Robert K. Merton's sociology of science, in particular his notion of the presence of a normative and a reward system in science (Merton, 1973, ch. 14). According to this perspective, a reference to a work (and thus a received citation of that work) is taken to serve, besides its instrumental function of pointing to work that might be of interest to the reader, a symbolic function because it registers the intellectual property of the acknowledged source by providing a small piece of peer recognition of the knowledge claim (Merton, 1988). Within this framework, then, citations are interpreted as indicators of merit of the cited work or the influence the work has had upon the relevant community of peers. Furthermore, the norms postulated by Merton predict that the authors cite work for scientifically relevant factors, i.e., the norm of universalism, and citing (or not) should not be influenced by the cited author's gender, ethnicity, or status in the scientific community. Of course, no one believes that norms and behaviors are perfectly correlated, as Zuckerman (1988) point out. However, proponents of this view hold that citations are a reasonable indicator of the influence of a scientific contribution and that by extension they signal

something about the merit of the work as defined by the scientific community.

A different interpretation of citations within the sociology of science is the so-called "constructivist view". Prominent examples of this view are presented in the work of Latour and Woolgar (1979) and Gilbert (1977). Here the focus is on rhetorical persuasions, and the bibliographic reference is seen as one important rhetorical device that an author has for persuading the reader of the merit of the scientific publication. Persuasion should not be understood in any normal sense of the word (it is trivially true that an author wants to persuade their readers that their work has some merit). Rather, the persuasion notion entails at least two types of disingenuous activities (White, 2004), persuasion by distortion (misrepresentation of the cited work on purpose) and persuasion by name-dropping (disproportionally citing authoritative authors or papers).

Variants and mixtures of the two perspectives are abundant, e.g., highly cited papers are conceptualized as concept symbols (Small, 1978), that the use of references is an important form of use of scientific information within the framework of documented science communication (Glänzel & Schoepflin, 1999) and/or that the rhetorical and the reward systems are concretely indistinguishable, both systems simultaneously motivating and constraining any given act of citing (Cozzens, 1989).

Several empirical tests of the strengths and predictive power of the two main theoretical interpretations of citations have been attempted. The general conclusion of such research favors a normative interpretation rather than a devious constructivist one in the explanation of observed citation patterns (Baldi, 1998; Judge, Cable, Colbert, & Rynes, 2007; Moed & Garfield, 2004; Riviera, 2014; Shadish, Tolliver, Gray, & Gupta, 1995; Stewart, 1983; Van Dalen & Henkens, 2001; P. Vinkler, 1998; Wang & Domas White, 1999; White, 2004).

Studies on citer motivation provide a more nuanced picture where the nature of the citing–cited relationship is scrutinized and individual references are manually classified according to their perceived function. Literature reviews of these studies are found in (Bornmann & Daniel, 2008; M. Liu, 1993; Small, 1982). Although these studies are difficult to compare because they do not use the same study designs or classification systems, they do suggest a rather high share (ranging from approximately from 10% to 50%) of perfunctory citations, i.e., the citing author cites another work without any obvious relevance to the citing author's immediate concerns. Possible explanations for these observations are that the reference list marks a

paper's "socio-cognitive location" and that the citing authors tend to ensure that important works are represented in the reference list. Even if a reference represents a cognitive influence, its expression in the text might be vague or implicit (Moed, 2005, ch. 16). Another explanation is based on Zipf's principle of least effort on which basis White (2001) hypothesized that perfunctory references are common simply because the effort involved in adding them is low (the same principle is said to explain why negative citations are relative rare, i.e., it is more of an effort to formulate an attack on an argument than to ignore it). These studies suggest that the topical content of the cited work might have a moderately constraining effect on its inclusion in the reference list of the citing work. A classic response to these studies – in an evaluative bibliometric context – is that the peculiarities that might be found in isolated reference lists does not weaken the normative interpretation to any significant degree as they do not shed much light on the collective effect of a community of citing authors (van Raan, 1998). Many idiosyncrasies associated with reference behavior can be expected – on statistical grounds – to play a minor role when analyzing large sets of documents and when the focus is on the cited side rather than the citing side.

If scientific and scholarly works can be assessed by the citations they receive, as suggested by the normative citation theory, it is natural to conduct criteria validation studies where citation indicators calculated for a unit of assessment are correlated with traditional peer review (the criterion). Peer review is often seen as an indispensible activity in most scientific areas because it enforces quality control and ensures trustworthiness in different scientific endeavors (Cronin, 2005). To proponents of peer review, equals (i.e., one's peers) working on the same or similar scientific problems are said to be in the best position to know whether quality standards have been met and a contribution to knowledge has been made (Eisenhart, 2002). The bulk of such validation studies report (usually in rank-order) correlations between citation-based evaluation and peer review grades in the range of 0.8–0.4 when the assessed unit is on the level of the department or research group (Aksnes & Taxt, 2004; Mahdi, D´Este, & Neely, 2008; O. Mryglod, R. Kenna, Y. Holovatch, & B. Berche, 2013; O. Mryglod, R. Kenna, Yu Holovatch, & B. Berche, 2013; Oppenheim, 1995, 1997; Rinia, van Leeuwen, van Vuren, & van Raan, 1998; Seng & Willett, 1995; Smith & Eysenck, 2002). Although such studies clearly demonstrate a statistical association between received citations and peer evaluation, the conclusions drawn must be somewhat limited. Firstly, the studies use different procedures for constructing citation indicators and examine different scientific areas, which makes generalization difficult. Secondly, it is not clear that peer review grading is a good criterion or "ground truth" against which citation-based assessment should be validated. The two methods might have quite different goals, e.g., peer

review of a department usually considers more parameters than the merit of past publications (Aksnes & Taxt, 2004), and thus one would expect a priori an upper bound for the correlation well below unity (Bornmann & Marx, 2013). Thirdly, the reliability of peer review is not necessarily very high (Allen, Jones, Dolby, Lynn, & Walport, 2009), and the chance factor in peer review outcomes can be quite substantial (Nederhof, 1988; Rothwell & Martyn, 2000) putting an upper limit for the correlations in these criterion-based evaluation studies at the level at which peer review correlates with itself.

While theoretical and empirical investigations into the appropriate conceptualization of citations is diverse, there is support for the idea – although with some reservations – that citations are a formalized account of information use and can thus be taken as an indicator of how the work is received among its peers (Glänzel, 2008). Thus, citations are often conceptualized as indicative of the actual influence a publication has on surrounding research activities at any given time, that is, its impact (Martin & Irvine, 1983). Essentially synonymous with impact is the notion that citations are indicative of the perceived utility of the scientific contribution. Attributes of knowledge claims are embedded in the formal research contributions, and these attributes influence the way the claims are received and will differ between research areas and over time. According to Cole (1992), these attributes are connected to the perceived utility of the scientific contribution. Utility has at least two components: the content of a document is useful if other researchers can build upon it or use it in their own work ("puzzle generating") and if it generates results that are expected ("puzzle solving"). Bibliographic references to earlier work can be seen as signals of perceived utility in either or both conceptualizations of the utility concept. The more peers cite a work, the greater influence the work tends to have on the surrounding research activities at a given time. However, research contributions can be greatly influential and rated highly on utility by peers but be virtually non-cited at a given time as an consequence of implicit citations – where the research contribution is decoupled from any reference to the source work (e.g., an instance of "the obliteration by incorporation phenomenon") – or by indirect citations where the reference is not given to the original research contribution but rather to a mediating work (e.g., an instance of "the palimpsestic syndrome") (MacRoberts & MacRoberts, 2010; McCain, 2014; Merton, 1973, p. 123; 1988). Thus, there is some inherent vagueness in the operationalization of impact and perceived utility by means of citations.

To complicate matters, it is unclear what constitutes quality of research and its formal representation. Presumably, the concept connects to a number of interacting factors such as originality, correctness, and intra- and extrascientific effects (Hemlin, 1993). However, quality of research is also a property that depends on the scientific problem area to which it belongs and thus only members working in this area can ultimately judge the quality of research (Gläser & Laudel, 2007). When citation counts are used in an evaluation, they are not used as a general measure of quality. Nevertheless, for it to be meaningful to use citation analysis to assess research, perceived utility and impact must be regarded to be at least *one aspect* of research quality. And if citations can be taken as the formalized use of information, we can study the judgment made by researchers active in the scientific problem area regarding the utility of different scientific contributions. To state that perceived utility is an aspect of the merit of scientific contributions is a rather moderate statement.

It should be noted that the above conceptualizations of citations are not applicable to all areas of scientific inquiry. Besides the technical coverage issue that a priori disqualifies universal application of citation-based performance exercises, one can argue that different research areas can be classified on a "hard–soft continuum". The research contribution on the softer end of the scale might be more open to greater interpretations and there might not be the same clear-cut criteria for establishing or refuting knowledge claims in the softer areas as in the harder. This results in different views about what constitutes a pertinent contribution and, by extension, affect the distribution of citations over documents (Hyland, 2004). Partly for such reasons, citations are argued by some to have a fundamentally different meaning in the softer spectra, and even if the technical limitations are alleviated, citation analysis as an evaluation tool would still be suspect in humanistic and related areas of scholarly inquiry (Hellqvist, 2010). For research areas that are a priori not suitable for citation analysis for evaluative purposes, other non-citation-based bibliometric approaches might be considered. These can be based on a researcher-driven quality classification of publishing channels like journals and publishing houses (Ahlgren, Colliander, & Persson, 2012; Schneider, 2009; Sivertsen, 2010).

## Pressing issues in the construction of citation indicators

While different theoretical perspectives on citations have been adopted, one can argue similarly to Zuckerman (1987) that the motives of citing authors and the consequences of these citations – which signal perceived utility or impact – are analytically distinct.

Assuming that not all citations are completely arbitrary and that not all citations given by researchers are biased in the same way, there are still two important problems connected to citation analysis for evaluative purposes that this thesis will try to address. Both of these problems are essentially independent of any sensible theoretical framing of what citations are indicative of. First, there is the question of how to enable comparisons between different documents. This is important because the raw numbers of received citations for documents that address disparate topics are largely incomparable. This follows from the fact that formal communication patterns differ with respect to such properties as the average length of reference lists, the proportion of recent references, the importance of different publication channels, coverage of the literature in the databases used for enumerating the citations, and the growth rate of the literature on a given subject or in a given research area. All these factors affects the probability that a document receives a citation regardless of its other qualities, and necessitates that the raw citation counts for a set of documents must be interpreted relative to some frame of reference. The traditional approach to handle this situation is to introduce the notion of *reference standards* or *reference sets*. These are sets of documents that should address similar research questions and, as a consequence, should be imbedded within similar formal communication contexts as the document attributed to the unit of assessment for which raw citation counts have been collected. Thus, "comparing 'like' with 'like' as far as possible" (Martin & Irvine, 1983, p. 61) is the basic principle for allowing fair application of citation-based evaluation by comparing the raw number of received citations to the documents in question with the distribution of citations in appropriate reference sets. How to operationalize an appropriate reference set for a document is, however, an open and vital question.

The second problem of evaluative citation analysis that will be addressed concerns how to handle the uncertainty connected to the process of attributing research publications to units of analysis. This requires aggregating the citations of these publications and quantifying the often-skewed distribution with some summary measure of, for example, the average citation impact. All empirical measures – whether based on bibliographic data or not – are associated with errors, and this should be taken into account when presenting bibliometric performance indicators.

# Background for the articles and general problem statements

A vital step in many different kinds of bibliometric investigations is the identification of documents that are similar in terms of their subject matter. The rationale for specific bibliometric investigations that depend on similarity estimates between documents can be radically different, from casting light on general insights into a contemporaneous state of knowledge (e.g., Small, 1999) to monitoring the scientific output from a research-producing unit and assessing its research performance on a detailed level (Noyons, Moed, & Luwel, 1999).

Formal scientific communications can be studied at different levels of aggregation depending on the specific goal of the study. However, there is no established nomenclature for classifying science at various levels. Concepts such as "disciplines", "fields", or "sub-fields" do not have any standard definitions and are used to imply different things by different authors and are often used synonymously (Ziman, 2000, ch. 8). That being said, an important concept is that of subject *specialties* or *problem areas*. These can be considered to be the largest homogeneous unit of science or scholarship in that each specialty has its own set of problems, a core group of researchers, and shared knowledge, vocabulary, and literature (Scharnhorst, Besselaar, & Börner, 2012). Because specialties play an important role in the creation and validation of new knowledge (Morris & Van der Veer Martens, 2008), it is of interest to study developments, discoveries, and conjectures generated within different specialties and to analyze the impact these contributions have on the progression of scientific and scholarly knowledge.

As far as bibliometrics is concerned, the underlying assumption is that research specialties can be fruitfully operationalized as evolving sets of documents of related subject matter (Lucio-Arias & Leydesdorff, 2009). Because publication and citation characteristics can vary substantially between specialties (Lillquist & Green, 2010), any inquiry concerning the number of citations received or the number of documents published by some unit must take this fact into consideration. The increasing focus on small units of assessment (i.e., below the level of country or university) in current research policy and citation-based evaluations increases the need for establishing appropriate frames of reference for contextualizing the raw citation impact of the documents that are attributed to such units (Rons, 2012).

A crucial question then, is how one can best identify documents that are related to the same subjects with the end goal of creating reference sets so that similar documents can be compared with each other.

## Bibliometric identification of subject-related documents

Subject similarity between scientific documents appraised by bibliometric methods is based on information present in the documents (or their surrogates in bibliographic databases) and meta-data attributed to the documents. Thus, there are essentially two sets of features or elements of the documents that can be explored in an effort to establish similarity relations, namely the cited references in the documents and the documents textual content. The latter case refers to the terminology used by the authors as well as potential indexing terms added by third-party subject specialists for enhancing information retrieval in bibliometric databases.

The use of cited references for establishing similarity relations is connected to the idea of citations as formalized accounts of information use. In particular, they are related to the notion that references cited in a document can be viewed as “subject terms” of that document and that the citing document has subject relevance to the ideas, methods, particular concepts, or hypothesis symbolized by the cited item (Garfield, 1964). Although this is the original raison d'être for bibliographic citation databases, this kind of first-order citation relationship might be of limited value in establishing subject similarity between documents, partly for reasons illuminated by studies of citer motivation and partly because documents published within the same time frame cannot have such relationships as a consequence of the inherent delay in working on a research problem and publishing its results.

Another approach to the identification of subject-similar documents by cited references is to consider higher-order citation relations between documents. That is to say, using cited references even though no direct citing–cited relationship necessarily exists. Bibliographic coupling (Kessler, 1963, 1965) is a concept that can be used to identify a subject similarity relationship between documents. Such a coupling occurs when two documents have one or more cited references in common. Similarly, the notion of co-citations (Small, 1973) states that two documents are co-cited if they are cited together by at least one other document. In both cases of these higher-order citation relationships, the more shared references a document pair have or, the more frequent the pair is co-cited, the higher the likelihood that they are related by subject matter.

Combinations of first and higher-order citation relationships between documents as a method for estimating topic similarity between documents are also possible. These methods either demand that several citation relations are present between a document pair – thus increasing the likelihood of a subject similarity connection – or that at least one among several potential citation relationships are present, thus increasing the coverage of document pairs for which there is estimated subject similarity (Persson, 2010; Small, 1997). There is evidence, however, that among these citation-based approaches bibliographic coupling outperforms other methods when the goal is to establish subject relatedness between documents and when high coverage is important (Boyack & Klavans, 2010).

The other type of feature found in the documents that can be exploited for identifying similarity relations is the textual content. Lexical coupling (Callon, Courtial, Turner, & Bauin, 1983) is present between documents when they share words, phrases, or index terms and thus have the potential to reveal subject similarity between the documents even if first or higher-order citation relations are absent for whatever reason. Lexical coupling can also provide additional evidence for the presence of topical similarity in cases where citation relations also exist. Although there is a high degree of codification in word usage in the scientific and technical literature (Leydesdroff, 1989), the likelihood that lexically coupled documents are topically similar increases when the coupling is based on highly specialized words and specific word classes such as nouns (Justeson & Katz, 1995). Variability in word usage that can decrease the effectiveness of lexical coupling, such as synonyms and word inflection, can potentially be reduced by converting words to their morphological root and by taking into consideration the correlation among words over the document set under study through techniques such as latent semantic analysis (Dumais, 2004).

Finally, one can envision some form of hybrid approach that combines both lexical coupling and citation relations in an effort to increase the likelihood of identifying subject-related documents (e.g., Janssens, Glänzel, & De Moor, 2008).

When document features are chosen as the basis for identifying topically similar documents, there is still the question of which specific similarity measure should be used to quantify the estimated similarity between document pairs. In principle, one could use the raw number of shared references or terms or the number co-citations as an estimate of similarity. However, using some form of transformation of the data, e.g., by relating the raw number of shared features in two documents to the total number of

features in the respective documents increases the accuracy of both citation (Boyack & Klavans, 2010) and lexical approaches (Klavans & Boyack, 2006).

Basically, similarity between two objects – documents, journals, etc. – can be measured in two essentially different ways. Either one focuses on the direct similarity between the two objects or one focuses on the way these objects relate to other objects in the population or dataset under study (Ahlgren, Jarneving, & Rousseau, 2003). These can be considered direct (or local) and indirect (or global) methods, respectively. Direct measures have been the standard approach to measure similarity between objects such as documents in bibliometric contexts. The main exception has been author co-citation studies (White & Griffith, 1982) where the objects are the authors' bodies of work and where indirect approaches are common. While many different direct similarity measures are available, many of them have a formal relationship to each other and the importance for subsequent analysis of the similarity data is not always dependent on the exact direct similarity measure that is used (Egghe, 2009). Nonetheless, when considering direct similarity measures there are arguments in favor (van Eck & Waltman, 2009) of probabilistic similarity measures (i.e., the deviation of observed overlap of document features from what would be expected if the features were independent) because these have properties that make them more suitable than set-theoretical similarity measures (i.e., the relative overlap of document features).

Although numerous studies have utilized bibliometric estimates of similarities between documents in exploratory studies to answer disparate empirical questions, the efforts to validate and detail the improvement in these approaches are rather sparse when compared to the validation efforts of applied approaches in other fields (Klavans, Boyack, & Small, 2012). In other words, it is important to establish the accuracy of different approaches for estimating the similarity between documents and not just to be content with the notion that different approaches present different insights into the phenomena being studied.

In particular, the notions of direct and indirect similarity are fundamentally different, and the usefulness of indirect similarity measures for identifying topically similar documents have not been sufficiently examined. Although Janssens (2007) observed a more distinct partitioning of documents when indirect similarity was used in combination with cluster analysis, other approaches to validation are needed if we want to examine whether this type of similarity calculation actually leads to increased accuracy when estimating subject similarity between documents.

## Normalizing raw citation impact with respect to subject matter

If we control for variations in reference behaviors and publication patterns in different specialties by relating the raw citation impact of a set of documents to other topically similar documents, we should be able to undertake meaningful investigations into the perceived utility of any set of documents. Akin to the notion of internal vs. external criteria for the assessment of research endeavors (Weinberg, 1963), indicators based on citation counts normalized in such a way correspond to the internal criteria insofar that we do not aim to differentiate between different specialties or scientific problem areas with respect to some notion of a hierarchy of importance. The aim of the assessment is to be able to identify documents whose contents are perceived at the given time of our investigation to be especially useful in the eyes of the researchers who are active in the specialty, or its associated specialties, in which the author(s) of the document is trying to make a contribution. The external criteria for the assessment of research concerns the question of why one should pursue one particular line of research in the first place, and this question is left to other types of investigations and justifications.

Surprisingly, the available toolset from bibliometric estimation of document–document similarity has not had much influence on the practice of contextualizing and normalizing citation counts in research evaluations. Instead, normalization of citation counts using reference sets based on the Subject Categories supplied by Thomson Reuters in the Web of Science has become, to use the words of Leydesdorff and Bornmann (2014, p. 1), "an established ("best") practice in evaluative bibliometrics". These Subject Categories are sets of journals as defined by the journal classification scheme used in Web of Science, which is arguably the de facto data source utilized for citation evaluation studies. The Subject Categories are, however, subjectively and heuristically defined and were originally created as a tool for information retrieval purposes (Pudovkin & Garfield, 2002). Their continuing importance and use in evaluative citation exercises are presumably of a rather pragmatic nature because they are usually considered "far from perfect, but [...] the only classification available" (Moed, Debruin, & van Leeuwen, 1995, p. 399).

These Subject Categories – around 220 in total not counting those related to the arts and humanities – are conceptualized as "fields of science" and their use as reference sets are based on the assumption that there is reasonable homogeneity with respect to reference behavior, communication patterns, and other factors within the sets that affect the probability of a document

being cited. However, several studies have shown a bias against certain research topics or specialties when Subject Categories are used for citation normalization because some documents, based on their subject, are embedded within quite different formal communication practices. Because of this fact, some documents naturally tend to receive more or fewer citations on average than documents in the same Subject Category that are addressing other topics. Such effects have been observed in the Library and Information Science category (Waltman, Yan, & van Eck, 2011), the Economics category (van Leeuwen & Medina, 2012), the Chemistry-related subject categories (Neuhaus & Daniel, 2009), and the medical-subject categories of Cardiac & cardiovascular systems, Clinical neurology, and Surgery (van Eck, Waltman, van Raan, Klautz, & Peul, 2013). For articles dealing with topics on Science and Technology Studies, it is even argued that using Subject Categories for citation normalization is simply impossible because such articles are spread out over a vast number of Subject Categories (Leydesdorff & Bornmann, 2014). In addition, there is no particular reason to doubt that problems of this kind are present in other subject specialties and in other Subject Categories. While vague assertions that subject heterogeneity within Subject Categories might be of less concern in practice for units of assessment on the macro level (at least at the university level) because different biases might cancel each other out (Schubert & Braun, 1996), no such reasoning seems reasonable when lower aggregations of documents are analyzed, i.e., at the institution or research group level.

Other subject-classification schemes for journals exist (e.g., Glänzel & Schubert, 2003; Rafols & Leydesdorff, 2009). From a more general perspective, though, the use of a journal or a set of journals as reference sets for citation normalization can be questioned. This is not only because a large diversity of articles on different subjects can be found within a single scientific journal (e.g., Boyack & Klavans, 2011; Glanzel, Schubert, & Czerwon, 1999), but also on the ground of Bradford's Law of Scattering (Bradford, 1934) and Garfield's Law of Concentration (Garfield, 1971). The first “law” relates to the tendency that articles on a given subject are found primarily in a small core set of journals, and the rest of the articles are spread out over other sets of journals that successively have to increase exponentially in the number of journals in order to contain the same number of articles on the subject as the core journal set. The second “law” asserts that for a given subject, many of the journals in these larger sets, with increasingly subject-irrelevant journals, are to a large extent part of the core set for some other subject areas. It is thus highly questionable to expect that journal sets in general will be homogeneous in terms of their subject matter (Leydesdorff & Bensman, 2006).

It should be noted that a completely different approach to normalization has been suggested that is based on the referencing behavior of the citing articles or citing journals (e.g., Leydesdorff & Opthof, 2010; Zitt & Small, 2008). The basic idea is to correct for differences in the length of the reference lists (the number of cited references) by weighting the received citations by some function of this length. The exact weighting can differ, and an overview of weighting tactics is given in (Waltman & van Eck, 2013a). The basic premise is the same, however, and the lengths of the reference lists (and the share of references that go to articles in the database within a given time period) in different research areas are taken as the main reason for different numbers of received citations observed between articles on disparate topics. Still, it is argued (Leydesdorff, Radicchi, Bornmann, Castellano, & de Nooy, 2013; Radicchi & Castellano, 2012) that this type of normalization does not remove citation biases between the literatures on different topics any more than traditional approaches. This is partly because the growth rate of the literature on a topic and unidirectional citations between, for example, applied and basic research literatures are not addressed by this type of normalization (Zitt & Small, 2008). However, conclusions that normalization based on some function of the length of the reference list are not better than traditional approaches are based on the use of a classification system, e.g., Subject Categories, in both the implementation and evaluation of the normalization approach (Sirtes, 2012; Waltman & van Eck, 2013b), and this might distort the results.

Perhaps a more radical point of view is given by Kostoff & Martinez (2005) who suggest that there might not exist a meaningful operationalization of concepts such as “fields” or “sub-fields” that is suitable for citation normalization. Rather, one should aim at comparing the citation count of different research articles with other articles that are as thematically (and temporally) similar as possible. Because there are relatively few articles in a given time period that are thematically very similar, Kostoff & Martinez (2005) argue that any metrics used to evaluate research should be based on this reality. One such approach entails a manually intensive approach of identifying those research articles most closely related to the articles whose citation counts are the subject of normalization and then using these identified articles as the basis for the normalization (Kostoff, 2002). Another approach involves using high-quality subject classification schemes that are available on the article level in specialized bibliographic databases. An example of this approach is the use of Medical Subject Headings descriptors for subject identification and citation normalization of medical research articles (Bornmann, Mutz, Neuhaus, & Daniel, 2008).

Although manually scrutinizing the published literature for documents that can be used for normalization purposes must be regarded as too unrealistic simply because of the workload involved, and although article-level subject classification schemes are only available for certain research areas, the general concept can still be developed. By using bibliometric methods for identifying topically similar documents, the citation impact of documents might be contextualized by relating them to other documents for which we have established a subject similarity connection. This avoids the problem with using journals as reference sets or on relying on the limited availability of article-level subject classification schemes. Potentially, one could also sidestep unclear notions of what a priori constitutes a reasonable aggregation level for the operationalization of the reference sets in the context of citation normalization by letting the notion of subject specialties dynamically define such reference sets based on empirical evidence.

## Uncertainty and robustness of citation indicators

Assuming that a reasonable solution to the problem of creating a meaningful frame of reference for calculating relative citation counts for a set of articles is attainable, there is the question of how to statistically address the level of uncertainty that is coupled with citation indicators.

There are errors in virtually all measurements. Some are non-sampling errors, which are errors that cannot be attributed to sampling fluctuations and might arise from many different sources. Sampling errors, on the other hand, are the difference between a population value and an estimate of that value that is due to the fact that only a particular sample of values are observed, and these are distinct from non-sampling errors (Dodge, 2003). Measures derived from bibliographic data are no different, although the first class of errors are much more prevalent in bibliometric research assessments because proper probability sampling is exceedingly rare in this context (Glänzel & Moed, 2012). Bibliometric indicators that summarize, for example, an empirical citation distribution into one or more values should, therefore, be accompanied by some information about how confident we are that a given indicator value is a good description of the underlying phenomena we want to say something about. Traditional frequentist statistical techniques aim to quantify the uncertainty that arises when generalizing from the sample to the entire population and deals with random errors that are generated from probability sampling or when we have random experimental designs. Because such situations are not common when units of assessment are subjected to evaluative citation analysis, other approaches should be explored.

Evaluative citation indicators are usually devoid of any estimates of uncertainty. Those that do use such estimates (e.g., Opthof & Leydesdorff, 2010; Schubert & Glänzel, 1983) usually adopt traditional inferential techniques that are designed to quantify sampling errors. However, because the basic premise of randomness for such approaches is clearly violated in most bibliometric studies, these estimates of uncertainty are ambiguous and hard to interpret at best and meaningless at worst. In a recent review, Schneider (2013) discussed the problem with classical inferential statistics and significance testing in the context of bibliometric citation evaluation and argued that the use of such tests does not provide any advantages in terms of deciding whether differences between citation indicators are important or not. Still, some defend the use of these procedures in bibliometric research assessment, for example, by pointing at other research areas such as psychology "where experiments are often based on convenience samples, and these tests are nevertheless carried out" (Bornmann & Leydesdorff, 2013, p. 1307) or that the observed bibliographic data for an assessed unit "might be thought of as being a sample from a larger super population that includes future cases as well" (Williams & Bornmann, 2014, p. 7). While the first argument is rather awkward, the second is also highly suspect because appeals to "super-populations" are generally considered invalid, especially in non-experimental social science settings (Berk, Wester, & Weiss, 1995; Schneider, 2014).

When a citation indicator for a unit, such as a department or university, is calculated, there can be counting errors and attribution errors – among other non-sampling errors – that improperly increase or decrease the indicator value. For some exercises there are estimates of the prevalence of such errors (N. C. Liu, Cheng, & Liu, 2005; van Raan, 2005), and in other cases at least some informed guesses can be made. Even if a study could be performed with a controlled "bottom-up" approach (van Leeuwen, 2005), where publications are collected from individual researchers' bodies of work and subjected to a verification round by the researchers themselves so that we can painstakingly prove virtually zero counting and attribution errors, it might still be of interest to supplement the indicators with some notion of uncertainty. By way of analogy, consider the case where we have a devised a measure of length that is of both high validity and reliability. If we measure the individuals in two distinct groups that do not represent probability samples from larger populations, and we summarize our measurement data with some indicator (e.g., the mean), we will probably find that the mean length differ. This would simply be a statement of fact and there would be no basis for proceeding with inferential statistics. Depending on the purpose of this exercise, the difference in mean length might, of course, be uninteresting. For example, if we randomly removed one or a few individuals

from the respective groups or if we randomly switched some of the individuals between the groups and recalculated the indicator again and came to the opposite conclusion then we might say that the original indicator values were not stable even if they were true and correct.

The notion of stability can be one way to augment the citation indicators and defend against over-interpretation both when a single unit is assessed in isolation or when several units are assessed, perhaps in a ranking context. We can operationalize stability by a computer-intensive resampling procedure (Lunneborg, 2000). Such a procedure can be conceptualized quite simply. Given the empirical citation distribution for a given unit, we calculate the indicator of choice repeatedly, but each time based on a large and random subset of the original citation distribution – i.e., sampling without replacement where the random sample is smaller than the original data.[1] This will give us a distribution of indicator values that tells us which values we would be likely to observe under small alterations of the original data. The form of this distribution is conditioned by both the original citation distribution and the indicator that is used.

The lower and upper percentiles from such a distribution can be used to create a stability interval for the calculated indicator. Because these intervals are based on percentiles from the subsample distribution, the intervals need not be symmetric around the observed indicator value. It is also not necessary to assume some particular functional form for the subsample distribution, e.g., by calculating the standard deviation and then relying on a Gaussian distribution.

The size of each subsample (e.g., 95% of the original data) could be guided by estimates of counting and attribution errors or otherwise be based on the investigator's threshold for what constitutes a reasonable notion of stability in any given context (e.g., taking into account the presence of potentially grossly erroneous reference values used in the normalization of the raw citation counts). Thus, a number of units of assessment could have different indicator values but overlapping stability intervals, and this would indicate that even though some unit's scores higher or lower on the utilized indicator than others, the differences between them are not stable and the observed differences might not be of particular interest. Or similarly, if one unit is followed over time holding the evaluation procedure constant, it is highly

[1] This is similar to bootstrapping (Efron & Tibshirani, 1994), which is a resample-based procedure to estimate standard errors. There are recent examples of the use of this approach in evaluative citation exercises (Chen, Jen, & Wu, 2014), but it assumes the availability of a proper probability sample and its rationale is to make traditional statistical inferences, albeit in a non-parametric manner.

probable that there will be *some* changes from one time period to another, but these changes might not be especially interesting if there are overlapping stability intervals. Conversely, non-overlapping intervals signal substantial differences in terms of stability, and this gives us more confidence when interpreting the differences in the calculated indicator values.

# Aim of the thesis

The purpose of this thesis is to contribute to the methodology at the intersection of relational and evaluative bibliometrics. Experimental investigations are presented that aim to address the question of how we can produce the most reliable estimates of the topic similarity between documents both automatically and by utilizing only information contained within the documents themselves. Results from these investigations are then explored within a context for creating frames of references in which raw citation counts can be contextualized for supporting internal criteria assessment of the degree to which scientific documents impact the advancement of the problem areas from which they originate and seek to influence.

To further provide a sound basis upon which one can draw informed conclusions with respect to observed levels of perceived utility of a document set, an approach that replaces the traditional notion of confidence intervals with that of resampling-based stability intervals is suggested and explored. This approach is motivated by the specific nature of bibliographic data and the data collection process utilized in citation evaluation studies. This latter concept is further introduced in the context of rankings – the part of citation-based studies that that usually gets the most attention – to highlight the instability that is inherent in many such exercises and to show how potentially incorrect conclusions might be drawn if notions such as stability of the derived citation indicators are ignored.

The above research questions are addressed in the four articles that make up this thesis:

I. Article 1 examines approaches for identifying the topical similarity between documents. The consequence of using both text and citation-based features derived from the documents and different methods for calculating similarity values are examined and validated with respect to a ground-truth classification of a test collection of documents that is supplied by a subject expert.

II. Article 2 is a follow-up to the tentative but promising results from Article 1. Using a large dataset, a specific method for deriving similarity estimates that take into account more global information than traditional similarity measures is shown to be more successful at identifying topical similarity between documents.

III. Article 3 draws from the insights of the preceding two articles to suggest and evaluate a method for deriving reference values for citation normalization that provides a more specific frame of reference than what is commonly used for assessing perceived utility by means of relative citation counts.

IV. Article 4 introduces the concept of stability in citation-based assessments and explores the ambiguity that follows from using different conventional reference sets at different levels of aggregation in citation-based evaluation studies.

# Results: Summary of the four articles

## ARTICLE I: Document–document similarity approaches and science mapping: Experimental comparison of five approaches

This paper experimentally compares five approaches, involving nine methods, for determining document–document similarity within the context of science mapping. We compare text-based approaches, the citation-based bibliographic coupling approach, and approaches that combine the two. Forty-three articles, published in the journal Information Retrieval, are used as test documents. We investigate how well the approaches agree with a ground-truth subject classification of the test documents when used in combination with a cluster analytic technique and with first-order and second-order types of similarities. The results show that it is possible to achieve a very good approximation of the classification by means of automatic grouping of articles. One text-only method and one combination method, with second-order similarities in both cases, give rise to cluster solutions that agree to a large extent with the classification.

A notable result is that the tested methods consistently perform better with second-order similarities, which are an instance of an indirect (i.e., global) similarity. For reasons connected to the relatively small size of the test collection and the fact that a validation methodology involving subject expert ground-truth classification is inherently somewhat subjective, more studies are needed on the similarity order issue.

## Article II: Experimental comparison of first and second-order similarities in a scientometric context

In this paper we use a large dataset to experimentally compare first-order with second-order similarities with respect to the overall quality of the partitions of the dataset where the partitions are obtained through a cluster analysis technique.

The dataset consists of 58,885 articles from the Abridged Index Medicus, which is a subset of the Medline database, and these articles are supplemented with cited references from Elsevier's Scopus database. We use the bibliographic coupling approach for the measurement of document–document similarity.

Because the issue of what constitutes the best number of clusters for a given dataset – irrespective of application – is an ill-posed and hard to solve issue, we have worked with a range of partitions – from fine-grained to coarse – and investigated if one of the similarity measures consistently performs better than the other.

The results show that second-order similarity consistently outperforms first-order similarity when the quality of a partition is defined in terms of the cluster's textual coherence.

## ARTICLE III: A novel approach to citation normalization: a similarity-based method for creating reference sets

In this paper, a similarity-oriented approach for deriving the reference values used in citation normalization is explored and contrasted with the dominant approach of utilizing database-defined journal sets as the basis for deriving such values. The study uses a subset consisting of 118,850 research articles covering a variety of research topics from Thomson Reuter's Web of Science.

Instead of trying to define disjoint reference sets, the similarity-oriented approach for deriving reference values defines as many reference sets as there are articles, and every article in the dataset has the potential to influence the reference set for a target article whose raw citation count is subject to normalization. The degree of influence is based on second-order similarity and utilizes a combination of bibliographic references and technical terminology. Thus an article's raw citation count is contrasted with topically similar documents within a fuzzy framework.

It is shown that reference values calculated by the similarity-oriented approach are considerably better at predicting the assessed article's citation count compared to the reference values given by the journal-set approach. This significantly reduces the variability in the observed citation distribution that stems from the variability in the article's addressed subject matter. Qualitative comparisons between the two approaches also suggest that the similarity-oriented approach makes the interpretation and meaning of a normalized citation count more straightforward and understandable. In contrast, in the subject-category approach the reference sets are highly subject-heterogeneous, and it can be difficult to interpret the derived normalized citation counts in this setting.

## ARTICLE IV: The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments

This paper presents a study on the effects of traditional, journal-based reference sets on the relative citation impact of 20 natural science departments at Stockholm University. The following three reference sets were used: the publishing journal, the Thomson Reuters Subject Categories, and the Essential Science Indicators fields. Citation impact was measured by the indicator item-oriented mean normalized citation rate and the proportion of top 5% publications. These indicators were calculated on the basis of three annual editions of Thomson Reuter's Web of Science.

We introduce a subsampling technique that can be applied when the data is neither randomly sampled nor randomly allocated (i.e., neither population nor causal inferences are feasible). Instead of talking about statistical significance (or lack thereof) we talk about stability, and a stable result is one that is not materially influenced by including or excluding specific documents that are attributed to a unit of assessment in the analysis.

We show that the ranking of a specific department, with respect to a given indicator, can differ not only within but also between normalization baselines. However, in many cases they do not differ in any substantial way as operationalized by the notion of stability. In light of the typically right-skewed nature of the underlying citation distribution, the subsample stability analysis has a clear merit in that it reveals the effect that a few documents might have on the indicator value and ward off over-interpretation by adding an interval to statements such as "unit A is cited x% above expectation", and this interval indicates how stable the observed indicator value is.

# Concluding discussion

The aim of this thesis was to combine relational and evaluative bibliometrics in an effort to enhance existing methods currently applied in citation-based research evaluations. A novel citation normalization methodology has been suggested that is based on a more direct notion of the idea of comparing "like with like". Together with the proposed approach for estimating the uncertainty that is inherent in citation evaluations, it is argued that these contributions have the potential to enable fairer citation-based evaluation exercises.

Regardless of whether a bibliographic coupling or lexical coupling approach is used, the results from Articles I and II together with other supporting validation studies (Cribbin, 2011) suggest that the second-order similarity method rather than the first-order method should be considered when estimating similarities between documents. Because the second-order approach, but not the first-order one, is able to determine that two documents are similar by finding that there are other documents such that the two documents are both directly similar to each of these other documents, the sensitivity connected to first-order similarity that comes primarily from the synonym problems in the case of lexical coupling and from a generalized notion of synonymy in the case of bibliographic coupling are reduced when second-order similarity is used. Put differently, because authors naturally use slightly different words when describing the same concepts and because they can draw on different samples from the literature when referring to relevant prior studies, traditional local document–document similarity measures based on text and cited references are more susceptible to neglecting significant similarity connections between documents than the suggested global measure. The larger amount of data involved in the global approach, which in essence supplements the local similarity estimate of two documents with information regarding their respective neighborhoods as defined by similar documents identified in the local step of the procedure, increases the likelihood of identifying topically similar documents.

The issue regarding the best approach to normalizing citation counts must be regarded as an ongoing research question. Article III uses a topic-level approach based on second-order similarity to open up a new perspective that shows promising advantages over more traditional journal-based normalizations. It is shown that traditional approaches to create reference sets from which relative citation indicators are derived adhere only weakly to the principle of "comparing like with like" and that the heterogeneous nature

of these reference sets calls into question the reasonability and the interpretability of such approaches. When the relative number of citations for an assessed unit is instead operationalized in terms of deviations from the expected number of citations received by documents that have been estimated to address similar research topics in an objective and quantifiable way, the contextualization of the raw citation counts is suggested to be more in line with the principle of "comparing like with like". It is shown that the number of citations received by a given document is highly dependent on which research topic is addressed, and this is not adequately handled by traditional normalization approaches.

If we take the scientific specialty as the largest homogeneous unit in science and consider that the validation and assessment of the utility of scientific contributions mainly take place within such units, it seems reasonable to construct a citation normalization approach that reflects this level of formal communication. Units in science larger than specialties, i.e., disciplines and fields, primarily perform infrastructure functions such as teaching, funding, and the institutional provision of libraries and laboratories (Morris & Van der Veer Martens, 2008), and it is generally not very meaningful to talk about research contributions being made to the advancement of knowledge in a discipline (Chubin, 1976). The size and fuzziness of specialties varies, but specialties tend to be relatively small units with shared knowledge, vocabulary, and archival literature because a researcher must be sufficiently familiar with the literature of their specialty to assess the novelty and/or plausibility of a knowledge claim in the specialty (Ziman, 2000, p. 182ff). By utilizing lexical coupling and bibliographic coupling, scientific contributions similar in subject matter that originate from and aspire to impact the same scientific problem area can be identified and act as sensible frames of reference for each other. To the extent that there is an underlying specialty for a research contribution, the proposed normalization approach can be considered to correspond reasonably well to this level.

Recently one can discern a tendency for more interest in constructing reference sets that are based on document–document similarity. In particular, work in progress presented in (Javier & Ludo, 2014) explores the creation of reference sets on the basis of clustering documents using direct citations as a measure of similarity between documents and then using the resulting clusters as reference sets. As these authors point out, the notions of scientific units such as specialties and fields do not have clear-cut boundaries; they overlap and their boundaries tend to be fuzzy. This is a fact that is not reflected by partitioning the scientific literature into disjoint sets by some clustering procedures. Furthermore, cluster solutions can be defined at many different levels of aggregation, and it is unclear which level

is most appropriate for the purpose of normalizing citation impact indicators. The approach to normalizing citations given in Article III might not give a definite solution to these hard to solve questions. However, this approach automatically incorporates the notion of fuzziness and levels of aggregation, firstly by allowing documents to participate in several reference sets with different weight and secondly by automatically adjusting the size of a reference set for a given publication based on the estimates of the number of available topic-relevant documents.

Compared to Article III, Article IV presents another type of evidence against using standard journal sets as a basis for normalization. Because using Subject Categories (more than 200) or Essential Science Indicators fields (only 22) does not lead to substantial differences in the interpretation of the citation indicators, this suggests that the heterogeneity with respect to the subject matter being addressed and to the citation traffic within the Subject Categories is so high that additional aggregation (adding additional subject-disparate documents to the mix) has only a minor effect on the derived reference values.

When differences between units of assessment are observed, such differences might not be substantial enough to draw any conclusions from them. The stability interval method introduced in Article IV provides a tool for augmenting the derived citation indicators with information regarding their robustness and helps to fend off over-interpretation of the data. Stability is operationalized by the degree to which one might draw different conclusions if small alterations are introduced into the document sets that are attributed to different units of assessment. Because there might be misattribution of documents to units, cases where the enumeration of citations are wrong or where a given reference value for a document is highly inappropriate, we should not present the citation indicators as being devoid of uncertainty. By using computer simulation, it is shown that one can study the effect that small alterations in the underlying data have on the derived indicator values and thus appreciate their stability or lack thereof. Indeed, small alterations that could be artifacts from the data collection and preparation step can have a significant influence on the picture painted by the citation indicator. Therefore, providing stability intervals around derived indicators prevents unfounded conclusions that otherwise could have unwanted policy implications.

It is worth pointing out that stability intervals are similar to traditional confidence intervals and significance testing in that both should not be used as mechanical tools for decision-making. The contexts and the specific goals of the citation performance exercises vary, and stability intervals are a tool to

support the interpretation of the data generated by these exercises. Given this caveat, it is clear that stability intervals have a straightforward practical value, and subsequent adoption of stability intervals (or some variation of them) by global bibliometric rankings such as The Leiden Ranking (Waltman et al., 2012) has greatly improved upon such endeavors.

As shown in Articles III and IV, a different basis for normalization can lead to a different interpretation of a unit's citation impact and expert-level input in the interpretation of citation indicators together with using other types of indicators of research performance is unavoidable if one wants to reach an acceptable level of validity. The notion of partially converging indicators (Martin, 1996; Péter Vinkler, 2011) is most important.

One of the greatest difficulties with validating any normalization procedure in citation analysis stems from the lack of any obvious gold standard to which a normalization approach can be compared. Future studies can, however, explore (1) how the generally applicable similarity-oriented approach compares to the normalization based on classification of documents into specialties that are available for certain research areas, e.g., by using Medical Subject Headings or the Physics and Astronomy Classification Scheme for article-level classification of biomedicine and physics literature, (2) if the proposed method for citation normalization increases the correlation with peer review assessments, and (3) if small-scale expert validations for selected subject specialties are possible. Taken together, such studies should increase (or decrease) our confidence in the proper way of handling citation normalization.

# References


Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Do metrics matter? *Nature, 465*(7300), 860-862.

Ahlgren, P., Colliander, C., & Persson, O. (2012). Field normalized citation rates, field normalized journal impact and Norwegian weights for allocation of university research funds. *Scientometrics, 92*(3), 767-780.

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology, 54*(6), 550-560.

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at a Norwegian university. *Research Evaluation, 13*(1), 33-41.

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for Landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS ONE, 4*(6).

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review, 63*(6), 829-846.

Berk, R. A., Wester, B., & Weiss, R. E. (1995). Statistical inference for apparent populations. *Sociological Methodology, 25*, 421-458.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*(1), 2-72.

Borner, K., Chen, C. M., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*, 179-255.

Bornmann, L. (2013). The problem of citation impact assessments for recent publication years in institutional evaluations. *Journal of Informetrics, 7*(3), 722-729.

Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80.

Bornmann, L., & Leydesdorff, L. (2013). Statistical tests and research assessments: A comment on Schneider (2012). *Journal of the American Society for Information Science and Technology, 64*(6), 1306-1308.

Bornmann, L., & Marx, W. (2013). How good is research really? - Measuring the citation impact of publications with percentiles increases correct assessments and fair comparisons. *Embo Reports, 14*(3), 226-230.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8*(1), 93-102.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404.

Boyack, K. W., & Klavans, R. (2011). *Multiple dimensions of journal specifity: Why journals can't be assigned to disciplines*. Paper presented at the The 13th Conference of the International Society for Scientometrics and Informetrics, Durban, South Africa.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 26*, 85-86.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191-235.

Chen, K.-M., Jen, T.-H., & Wu, M. (2014). Estimating the accuracies of journal impact factor through bootstrap. *Journal of Informetrics, 8*(1), 181-196.

Chubin, D. E. (1976). The conceptualization of scientific specialties. *The Sociological Quarterly, 17*(4), 448-476.

Cole, S. (1992). *Making science: Between nature and society*. Cambridge, Mass.: Harvard University Press.

Cozzens, E. S. (1989). What do citations count? The rhetoric-first model. *Scientometrics, 15*(5-6), 437-447.

Cribbin, T. (2011). Discovering latent topical structure by second-order similarity. *Journal of the American Society for Information Science and Technology, 62*(6), 1188-1207.

Cronin, B. (2005). *The hand of science : Academic writing and its rewards*. Lanham, Md: Scarecrow Press.

Dodge, Y. (2003). *The Oxford dictionary of statistical terms*. Oxford: Oxford University Press.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188-230.

Efron, B., & Tibshirani, R. (1994). *An introduction to the bootstrap*. New York: Chapman & Hall.

Egghe, L. (2009). New relations between similarity measures for vectors based on vector norms. *Journal of the American Society for Information Science and Technology, 60*(2), 232-239.

Eisenhart, M. (2002). The paradox of peer review: Admitting too much or allowing too little? *Research in Science Education, 32*(2), 241-255.

Garfield, E. (1964). The citation index - A new dimension in indexing. *Science, 178*, 471-479.

Garfield, E. (1971). The mystery of the transposed journal lists: Wherein Bradford's Law of Scattering is generalized according to Garfield's Law of Concentration. *Current Contents, 3*(33), 5-8.

Gilbert, N. (1977). Referencing as persuasion. *Social Studies of Science, 7*(1), 113-122.

Glänzel, W. (2008). Seven myths in bibliometrics about facts and fiction in quantitative science studies. *Collnet Journal of Scientometrics and Information Management, 2*(1), 9-17.

Glänzel, W., & Moed, F. H. (2012). Opinion paper: Thoughts and facts on bibliometric indicators. *Scientometrics*.

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management, 35*(1), 31-44.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3), 357-367.

Glanzel, W., Schubert, A., & Czerwon, H. J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics, 44*(3), 427-439.

Gläser, J., & Laudel, G. (2007). The social construction of bibliometric evaluations. In R. Whitley & J. Gläser (Eds.), *The Changing Governance of the Sciences* (Vol. 26, pp. 101-123): Springer Netherlands.

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology, 61*(2), 310-318.

Hemlin, S. (1993). Scientific quality in the eyes of the scientist. A questionnaire study. *Scientometrics, 27*(1), 3-18.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*(2), 251-261.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor, Michigan: University of Michigan press.

Janssens, F. (2007). *Clustering of scientific fields by integrating text mining and bibliometrics*. (Unpublished doctoral dissertation), Katholieke Universiteit, Leuven.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics, 75*(3), 607-631.

Javier, R.-C., & Ludo, W. (2014). Field-normalized citation impact indicators using algorithmically constructed classification systems of science *UC3M Working papers (14-3)*. http://hdl.handle.net/10016/18385.

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited - Article, author, or journal? *Academy of Management Journal, 50*(3), 491-506.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural language engineering, 1*(1), 9-27.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10-25.

Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation, 16*(3), 223-233.

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology, 57*(2), 251-263.

Klavans, R., Boyack, K. W., & Small, H. (2012). *Indicators and precursors of 'hot science'.* Paper presented at the 17th International Conference on Science and Technology Indicators, Montreal, Canada (2012).

Kostoff, R. N. (2002). Citation analysis of research performer quality. *Scientometrics, 53*(1), 49-71.

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

Latour, B., & Woolgar, S. (1979). *Laboratory Life*. Princeton, New Jersy: Princeton University Press.

Leydesdorff, L., & Bensman, S. (2006). Classification and powerlaws: The logarithmic transformation. *Journal of the American Society for Information Science and Technology, 57*(11), 1470-1486.

Leydesdorff, L., & Bornmann, L. (2014). The operationalization of "fields" as WoS Subject Categories (WCs) in evaluative bibliometrics: The cases of "Library and Information Science" and "Science & Technology Studies". *ArXiv e-prints, 1407*, 7849. http://arxiv.org/abs/1407.7849

Leydesdorff, L., & Opthof, T. (2010). Normalization at the field level: Fractional counting of citations. *Journal of Informetrics, 4*(4), 644-646.

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.

Leydesdroff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*(4), 209-223.

Lillquist, E., & Green, S. (2010). The discipline dependence of citation statistics. *Scientometrics, 84*(3), 749-762.

Liu, M. (1993). Progress in documentation - The complexities of citation practice: A review of citation studies. *Journal of Documentation, 49*, 370-408.

Liu, N. C., Cheng, Y., & Liu, L. (2005). Academic ranking of world universities using scientometrics - A comment to the "Fatal Attraction". *Scientometrics, 64*(1), 101-109.

Lucio-Arias, D., & Leydesdorff, L. (2009). An indicator of research front activity: Measuring intellectual organization as uncertainty reduction in document sets. *Journal of the American Society for Information Science and Technology, 60*(12), 2488-2498.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. Australia: Duxbury.

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology, 61*(1), 1-12.

Mahdi, S., D´Este, P., & Neely, A. (2008). *Citation counts: Are they good predictors of RAE scores?* Bedford, UK: Advanced Institute of Management Research Retrieved from http://ssrn.com/abstract=1154053.

Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics, 36*(3), 343-362.

Martin, B. R., & Irvine, J. (1983). Assessing basic research - Some partial indicators of scientific progress in radio astronomy. *Research Policy, 12*(2), 61-91.

McCain, K. (2014). Assessing obliteration by incorporation in a full-text database: JSTOR, Economics, and the concept of "bounded rationality". *Scientometrics*, 1-15.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: Univ. of Chicago Pr.

Merton, R. K. (1988). The Matthew Effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis, 79*(4), 606-623.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Moed, H. F., Debruin, R. E., & van Leeuwen, T. (1995). New bibliometric tools for the assessment of national research performance - Database Description, overview of indicators and first applications. *Scientometrics, 33*(3), 381-422.

Moed, H. F., & Garfield, E. (2004). In basic science the percentage of 'authoritative' references decreases as bibliographies become shorter. *Scientometrics, 60*(3), 295-303.

Morris, S. A., & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology, 42*, 213-295.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013). Absolute and specific measures of research group excellence. *Scientometrics, 95*, 115-127.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. (2013). Comparison of a citation-based indicator and peer review for absolute and specific measures of research-group excellence. *Scientometrics, 97*(3), 767-777.

Narin, F. (1976). *Evaluative bibliometrics : The use of publication and citation analysis in the evaluation of scientific activity*. Washington, D.C.: Computer horizons, Inc.

Nederhof, A. J. (1988). The validity and reliability of evaluation of scholarly performance. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 193-228). Amsterdam: Elsevier.

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.

Noyons, E. C. M., Moed, H. F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science, 50*(2), 115-131.

Oppenheim, C. (1995). The correlation between citation counts and the 1992 research assessment exercise ratings for British library and information science university departments. *Journal of Documentation, 51*(1), 18-27.

Oppenheim, C. (1997). The correlation between citation counts and the 1992 research assessment exercise ratings for British research in genetics, anatomy and archaeology. *Journal of Documentation, 53*(5), 447-487.

Opthof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics, 4*(3), 423-430.

Persson, O. (2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics, 4*(3), 415-422.

Pudovkin, A. I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. *Journal of the American Society for Information Science and Technology, 53*(13), 1113-1119.

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.

Rafols, I., & Leydesdorff, L. (2009). Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. *Journal of the American Society for Information Science and Technology, 60*(9), 1823-1835.

Rinia, E. J., van Leeuwen, T., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands. *Research Policy, 27*(1), 95-107.

Riviera, E. (2014). Testing the strength of the normative approach in citation theory through relational bibliometrics: the case of Italian sociology. *Journal of the American Society for Information Science*. doi: 10.1002/asi.23248

Rons, N. (2012). Partition-based Field Normalization: An approach to highly specialized publication records. *Journal of Informetrics, 6*(1), 1-10.

Rothwell, P. M., & Martyn, C. N. (2000). Reproducibility of peer review in clinical neuroscience: Is agreement between reviewers any greater than would be expected by chance alone? *Brain, 123*(9), 1964-1969.

Scharnhorst, A., Besselaar, P., & Börner, K. (2012). *Models of Science Dynamics*. Berlin, Heidelberg: Springer.

Schneider, J. W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. *European Political Science, 8*(3), 364-378.

Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics, 7*(1), 50-62.

Schneider, J. W. (2014). Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics*. doi: 10.1007/s11192-014-1251-5

Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics, 36*(3), 311-324.

Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics, 5*(1), 59-73.

Seng, L. B., & Willett, P. (1995). The citedness of publications by United Kingdom library schools. *Journal of Information Science, 21*(1), 68-71.

Shadish, W. R., Tolliver, D., Gray, M., & Gupta, S. K. S. (1995). Author judgements about works they cite: Three studies from psychology journals. *Social Studies of Science, 25*(3), 477-498.

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.

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. *ISSI Newsletter, 6*(1), 22-28.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics, 91*(2), 567-575.

Small, H. (1973). Co-citation in the scientific literature: A New measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265-269.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327-340.

Small, H. (1982). Citation context analysis. In B. Dervin, G. J. Hanneman & M. J. Voigt (Eds.), *Progress in Communication Sciences* (Vol. 3, pp. 287-310). Norwood, N.J.: Ablex.

Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics, 38*(2), 275-293.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799-813.

Smith, A., & Eysenck, M. (2002). *The Correlation between RAE ratings and citation counts in psychology*. Department of Psychology, Royal Holloway. University of London. London. Retrieved from http://cogprints.org/2749/

Stewart, J. A. (1983). Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces, 62*(1), 166-189.

Van Dalen, H. P., & Henkens, K. (2001). What makes a scientific article influential? The case of demographers. *Scientometrics, 50*(3), 455-482.

van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology, 60*(8), 1635-1651.

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. (2005). Descriptive versus evaluative bibliometrics. In H. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative*

*Science and Technology Research* (pp. 373-388). Netherlands: Springer.

van Leeuwen, T., & 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. (1998). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics, 43*(1), 129-139.

van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133-143.

Vinkler, P. (1998). Comparative investigation of frequency and strength of motives toward referencing, the reference threshold model. *Scientometrics, 43*(1), 107-127.

Vinkler, P. (2011). Application of the distribution of citations among publications in scientometric evaluations. *Journal of the American Society for Information Science and Technology, 62*(10), 1963-1978.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., . . . Wouters, P. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419-2432.

Waltman, L., & van Eck, N. J. (2013a). Source normalized indicators of citation impact: A overview of different approaches and an empirical comparison. *Scientometrics, 96*(3), 699-716.

Waltman, L., & van Eck, N. J. (2013b). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics, 7*(4), 833-849.

Waltman, L., Yan, E., & van Eck, N. J. (2011). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics, 89*(1), 301-314.

Wang, P., & Domas White, M. (1999). A cognitive model of document use during a research project. Study II. Decisions at the reading and

citing stages. *Journal of the American Society for Information Science, 50*(2), 98-114.

Weinberg, A. M. (1963). Criteria for scientific choice. *Minerva, 1*(2), 159-171.

White, H. D. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology, 52*(2), 87-108.

White, H. D. (2004). Reward, persuasion, and the Sokal Hoax: A study in citation identities. *Scientometrics, 60*(1), 93-120.

White, H. D., & Griffith, B. C. (1982). Authors and markers of intellectual space: Co-citation studies of science, technology and society. *Journal of Documentation, 38*, 255-272.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Science and Technology, 32*, 99-168.

Williams, R., & Bornmann, L. (2014). Sampling issues in bibliometric analysis. *ArXiv e-prints, 1401*, 2254. http://arxiv.org/abs/1401.2254

Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport, Conn. : Libraries Unlimited.

Ziman, J. M. (2000). *Real science : what it is, and what it means*. Cambridge: Cambridge University Press.

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.

Zuckerman, H. (1987). Citation analysis and the complex problem of intellectual influence. *Scientometrics, 12*(5-6), 329-338.

Zuckerman, H. (1988). The Sociology of Science. In N. J. Smelser (Ed.), *Handbook of Sociology* (pp. 511-574). California: Sage Publications.