



Universiteit
Leiden
The Netherlands

Bibliometric uncertainty: data quality, reliability, and validity

Donner, P.

Citation

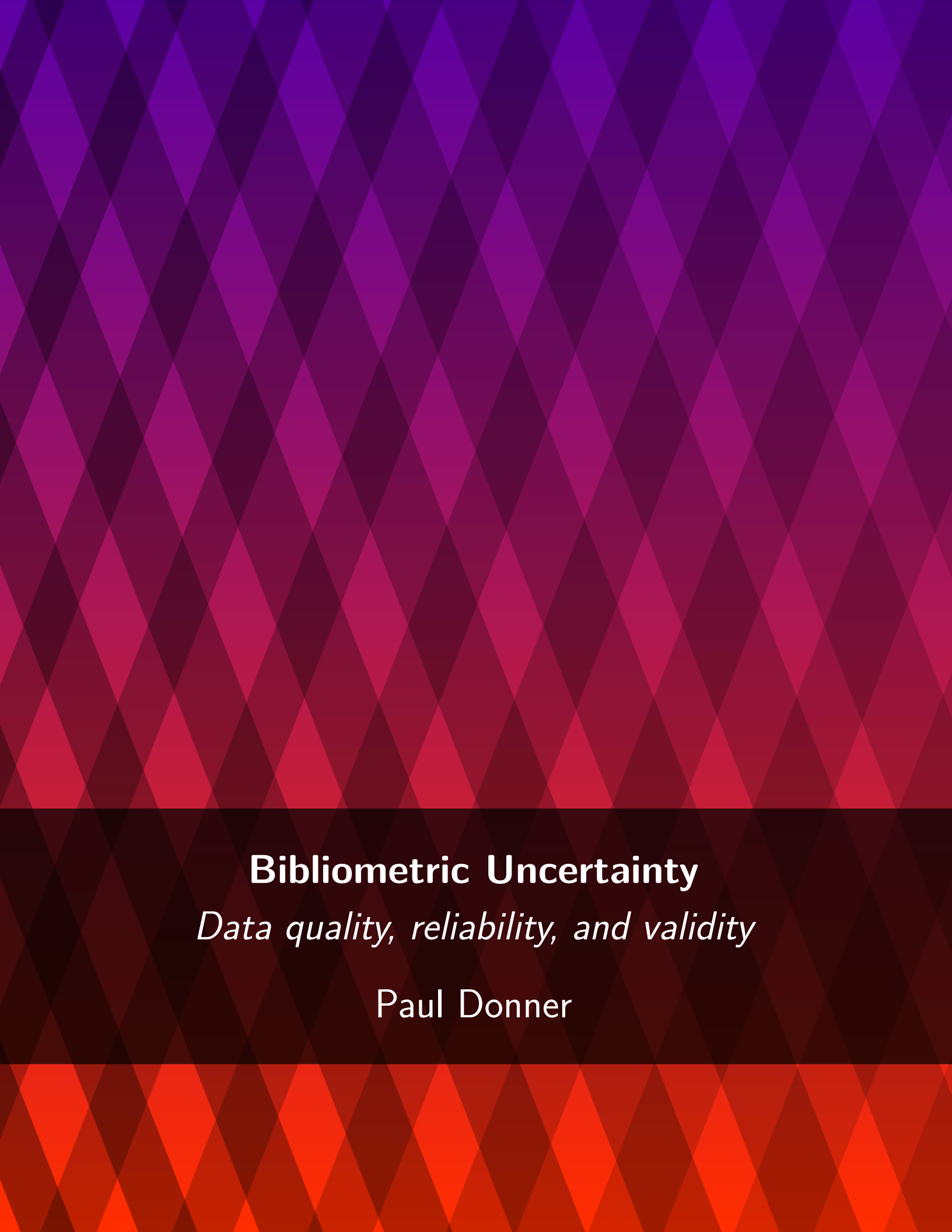
Donner, P. (2026, October 1). *Bibliometric uncertainty: data quality, reliability, and validity*. Retrieved from <https://hdl.handle.net/1887/4310427>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4310427>

Note: To cite this publication please use the final published version (if applicable).



Bibliometric Uncertainty

Data quality, reliability, and validity

Paul Donner

Bibliometric uncertainty: Data quality, reliability, and validity

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus
prof.dr. Sarah de Rijcke,
volgens besluit van het college voor promoties
te verdedigen op donderdag 1 oktober 2026
klokke 11.30 uur
door

Paul Donner
geboren te Berlijn
in 1985

Promotor

Prof.dr. Paul Wouters

Co-promotor

Dr. Nees Jan van Eck

Promotiecommissie

Prof.dr. Ludo Waltman (Wetenschappelijk Directeur Instituut CWTS/voorzitter)

Prof.dr. Cinzia Daraio (Sapienza Università di Roma)

Prof.dr. Tim Engels (Universiteit Antwerpen)

Prof.dr. Thed N. van Leeuwen

Prof.dr. Jesper W. Schneider (Aarhus Universitet)

Contents

Acknowledgments	vi
1 Introduction	1
1.1 The socio-political environment of bibliometrics	1
1.2 Terminology	5
1.3 Conceptions of data quality	6
1.3.1 Definitions	6
1.3.2 Demands of data quality in basic versus applied bibliometrics	7
1.3.3 Dimensions of data quality	8
1.3.4 Other distinctions of data quality in bibliometrics	10
1.4 Data sources	12
1.5 Indicator use and the importance of bibliometric data	15
1.5.1 Non-opposition of qualitative and quantitative research assessment	16
1.5.2 Beginnings of bibliometrics	19
1.5.3 The present role of bibliometrics	21
1.6 Bibliometric measurement uncertainty as a research subject	23
1.6.1 Elements of uncertainty	24
1.6.2 A panorama of sources of uncertainty in bibliometrics	29
1.7 Organization of the dissertation	38
2 Document type assignment accuracy in the journal citation index data of Web of Science	42
2.1 Introduction	42
2.2 Methods	46
2.2.1 Data and Sampling	46

2.2.2	Data collection	48
2.2.3	Data analysis	49
2.3	Results and discussion	51
2.4	Conclusion	61
	Appendix	62
	A: Matching of WoS and Scopus records	62
	B: Distribution of publication years in the WoS sample	63
	C: Equality of population estimates of Precision and Recall	63
3	Effect of publication month on citation impact	65
3.1	Introduction	65
3.2	Related work	67
3.3	Data and methods	69
3.3.1	Data sets	69
3.3.2	Missing month data	70
3.3.3	Indicators and statistical methods	71
3.4	Results	72
3.4.1	Missing month data estimation and validation	72
3.4.2	Effect of publication month on citations at the level of month-cohort groups of papers	73
3.4.3	The size of the publication month effect in different disci- plines and its relation to citation dynamics indicators	77
3.4.4	Effect of publication month at the level of individual pub- lications	79
3.4.5	Effect of online publication month	80
3.4.6	Month level precision citation windows	82
3.4.7	Publication month bias in citations in practice – the influ- ence on institution impact scores	85

3.5	Discussion	89
3.5.1	Limitations	89
3.5.2	Discussion of findings	89
	Appendix	91
	Author self-citation	91
	Month bias in different disciplines	93
4	Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems	94
4.1	Introduction	94
4.2	Related work	96
4.2.1	Rule-based approaches	96
4.2.2	Machine learning approaches	97
4.3	Systems, data, and methods	98
4.3.1	Institution disambiguation systems	98
4.3.2	Data	100
4.3.3	Methods	103
4.4	Results	105
4.4.1	WoS Organizations Enhanced	106
4.4.2	WoS Institution Name Search	108
4.4.3	Scopus AFID	111
4.5	Discussion	115
	Appendix A: Evaluation of the KB disambiguation system for German research institutions	118
5	Data inaccuracy quantification and uncertainty propagation for bibliometric indicators	121
5.1	Introduction	121

5.2	Related work	124
5.2.1	Uncertainty in scientometrics not related to data errors .	124
5.2.2	Accuracy and completeness of citation links in bibliographic databases	126
5.3	Missing links: Empirical study of citation error distribution . . .	132
5.3.1	Methods and data	132
5.3.2	Results	136
5.4	Propagation of uncertainty from data to bibliometric indicators – a Bayesian regression approach	138
5.4.1	Incorporating information about uncertainty into statistical models	138
5.4.2	A simulation exercise of bibliometric uncertainty propagation – Simulating error-free data	140
5.4.3	A real-world example application	147
5.5	Discussion	151
5.5.1	Summary	151
5.5.2	Limitations and future work	153
5.5.3	Conclusion	155
	Appendix	156
	5.5.4 Simulating error-affected data	156
6	General conclusion	162
6.1	Summary of findings	162
6.2	Consequences for bibliometric studies and bibliometric data sources	167
6.2.1	General implications	167
6.2.2	Research assessment reform	168
6.2.3	Improvements in data quality	169
6.2.4	Open data, open methodology, and transparency about uncertainty	170

6.2.5	Methodological research in bibliometrics	172
6.2.6	Adoption of explicit uncertainty reporting	173
6.3	Limitations	175
6.4	Further research directions	178
References		181
English language summary		203
Nederlandstalige samenvatting		204
Curriculum Vitae		205

Acknowledgments

I would like to thank my doctoral advisers Paul Wouters and Nees Jan van Eck for their very generous and incredibly helpful support. Throughout the course of the whole project they have both given me freedom to make my own scientific choices where it mattered most and offered and given the help, advice, and motivation that was needed at the right moment. My goal in this project was to do research that satisfies my personal standards, which I would consider quite stringent. Nees Jan and Paul have subtly communicated that this is a necessary but not a sufficient precondition for good research.

Part of the work for this dissertation has benefited from allocation of working time at my employer DZHW for this doctoral research project and from the access to research infrastructure at DZHW. For this, I would like to express my gratitude. In this regard, I would like to thank in particular Sybille Hinze and Stephan Stahlschmidt for their continued support. I would be remiss not to acknowledge my debt to my university lecturers Michael Heinz and Frank Havemann, who initially kindled my interest in the research topic of bibliometrics and scientometrics.

Much of the original research used proprietary data made available in the infrastructure of the Kompetenznetzwerk Bibliometrie, previously Kompetenzzentrum Bibliometrie, funded by the German Federal Ministry of Research, Technology, and Space.

1 Introduction

Any large-scale research assessment exercise, as well as any mapping of research fields for sociological or science policy purposes, is heavily conditioned by the quality of the information retrieval and data-cleansing routines applied to the source database. (De Bellis, 2009, p. xxii)

This dissertation is a methodological contribution to bibliometric research building on empirical investigations. In the four original research chapters, I critically examine important aspects of the fundamental validity of the conventional approach to core bibliometric methodology – publication and citation analysis. I introduce newly developed methods to overcome existing shortcomings and to improve the reliability and validity of citation analysis, and to express bibliometric indicator uncertainties explicitly that follow from imperfect data quality.

To get to the question of how to better handle data errors and inaccuracies in bibliometrics, I will begin this dissertation with the more fundamental question of why bibliometrics and metrics-based research assessment are done at all and why they are done in the particular way they are nowadays done. To this end, I need to sketch the sociopolitical environment in which bibliometrics developed and which it continues to be part of.

1.1 The socio-political environment of bibliometrics

Scientific research has over the course of the twentieth century become such an integral part of modern societies that its current organization, in which most basic research is carried out by fully or largely publicly supported professionals in research universities and dedicated research organizations, might be taken for granted. In financing academic basic research on a grand scale, the financier, usually a state, would do well to find some means to assure that its subsidies are not squandered or used too inefficiently. If that were to happen, public opinion may turn against, not only the subsidized research system, but also against that part of the state that has allowed things to come so far and which the public assumes to bear responsibility for the diligent and efficient use of state funds. All state funds are, after all, resources generated by the public through their economic activity but requisitioned by the state through taxation

and thereby withheld from private consumption and saving and directed towards consumption as directed by the state. A state thus finds itself in a position where it must exercise oversight over the spending of its dispensations to academic research lest it risks drawing the ire of the public on itself for neglecting to do so if things go awry (Elzinga, 2012).

In pursuit of this objective, states face resistance by the organized profession of researchers which insists on a large degree of autonomy in what it considers matters internal to the profession, such as personnel decision-making, priority setting of research agendas, selection of projects for funding and demarcation of science from non-science. The profession rejects outside steering as undue interference on the grounds that such processes require a degree of specialized knowledge and tacit experience that non-professionals do not possess, a claim that Bimber & Guston (1995) (p. 557) refer to as the Platonic exceptionalism of science. Although existing for much longer, over the course of the past decades, the family of judgement and decision-making procedures which can be subsumed under the label of formal peer review has attained a central position in practical but also in ideological terms due to its embodiment of the terms of the discourse as set out by the research profession: scientific and scholarly judgement can only be legitimate if exercised by qualified peer researchers. Outsiders, such as politicians and bureaucrats, are held at arm's length by the profession by arguing that if science is to achieve its goals, including objectives commissioned by outside societal consensus, the most rational way to do this is by relying on specialist scientists' judgments, by peer review.

In contrast, applied bibliometrics, has partially taken on the function of a technology and a means of state legitimization¹. It has that function, even if it perhaps was not explicitly and consciously developed for it and if perhaps some practitioners find this characterisation distasteful. It is hence neither surprising that bibliometrics in its incarnation as a basic science hardly studies non-academic, industrial research, nor that bibliometrics in its application in research evaluation primarily finds use in the assessment of publicly funded research rather than privately funded research, be it private enterprise or private philanthropy, precisely because these types of research financiers are not

¹Godin (2005) put it this way: "But most often, governments constructed statistics to serve *symbolic, ideological* and *political* ends. Statistics certify the user's credibility and seriousness. It objectifies political choices, that is, eliminates arbitrariness and inherent subjectivity in political action by portraying decisions as transparent and lending legitimacy to political discourse. S&T statistics is no exception." (p. 5, italics in original)

accountable to the public. It is not easy to imagine what bibliometrics would look like in the absence of large-scale state funding in research-based knowledge production, but it is obvious that it would appear very different. This does not mean that bibliometrics has no *raison d'être* independent of state demand for research evaluation and oversight, only that this demand has historically shaped its development (Godin, 2005; Wouters, 1999, p. 131).

One can find acknowledgment of this connection between knowledge production and the use of bibliometrics for research evaluation in the professional bibliometric literature. For instance, Vinkler (2010) stated that “The aim of scientometrics is to reveal characteristics of scientometric phenomena and processes in scientific research for more efficient management of science” (p. 1). De Bellis (2009) wrote:

From a science policy standpoint, it [the recognition of scientific quality] is also a useful guide in the allocation of the financial and human resources necessary for the machinery of science to work properly. (p. 181)

The peer review crisis triggered off the search for alternative or complementary evaluation methods more detached from subjective judgments. So, at least from the 1970s, the use of citation analysis to produce indicators of scientific performance in terms of citation scores began to receive unprecedented attention from politicians and science managers, clearly attracted by the several virtues of a numerical index: convenient, quickly understood, easily applied, and, above all, easy to calculate thanks to ISI databases. (p. 183)

An OECD working paper on bibliometric indicators stated:

A change took place in the 1970s: science was no longer seen as a venture in which society could invest generously and without limit. [...] science and technology were expensive, but investment in research did not automatically make it possible to solve environmental problems or social problems [...] This led to concern over the profitability of basic research in particular, and researchers were increasingly perceived as “producers of science” who had to account for the funds

they received. [...] The new aim was to produce value added that conserved natural resources and created less pollution, and to create a more efficient research system that made better use of existing intelligence. It was in this context that the evaluation of scientific research came into its own. [...] methods had to be found to quantify, compile and compare indicators. The establishment of a measurement of science(s) became inevitable. This shift cleared the way for the analysis of science and technology and favoured the advent of bibliometricians in science policy. (Okubo, 1997, pp. 11–12)

The emergence of bibliometric indicators – as primarily a science policy instrument – is traced historically by Godin (2005), chapter 6, to the publication of the first *Science Indicators* report of the US National Science Foundation in 1973, developed by its National Science Board. It was widely read and discussed and sparked the publication of various national and international report series that followed its model. The first report included publication and citation statistics calculated from the Science Citation Indexes which were prepared by Computer Horizons, Inc.

While bibliometric indicators are presently used extensively also within the semi-autonomous research system (see section 1.5.3) they were up to a point historically primarily developed as a tool for monitoring developments in the research systems from outside. Bibliometrics could serve as a tool to counterbalance a progressing emancipation of a societal subsystem. Complete independence and unchecked self-administration of science is not in the interest of states, which would continue to bear the responsibility of financing increasing demands while losing the wherewithal to verify the prudent spending of funds. The logic of rationality that bibliometrics derives from (purported) objectivity and quantification makes it an ideal tool of observation, steering, and decision-making at arm's length, from outside the system itself. A further advantageous feature of bibliometrics as such a tool is that it works, at a first approximation, without directly interfering with the day-to-day conduct of scientific research, that is, it does not infringe on the prerogatives of the profession.

We shall take the large-scale, state-mediated public funding of scientific research and the role that bibliometrics has assumed in this environment as given for this dissertation². We shall ask, *given* financial state support for academic research,

²Suffice it say that economists are by no means in unanimous agreement that government

given the need for accountability consequently demanded by the public bearing the costs of these expenses, how can bibliometrics contribute meaningfully to such oversight and how does the quality of bibliometric data matter for such a purpose.

1.2 Terminology

Bibliometrics is one of several methodological approaches to science studies or scientometrics, that is, the scientific investigation of the socio-cognitive process of scientific activity itself.

The term ‘bibliometrics’ was popularized by Alan Pritchard in 1969 (Pritchard, 1969) although its French version, *bibliométrie*, was used earlier by Paul Otlet (Rousseau et al., 2018). Pritchard defined bibliometrics as “the application of mathematics and statistical methods to books and other media of communication.” Historical overviews of bibliometrics, corroborating the close co-development with large-scale involvement of states in supporting basic research can be found in Van Raan (2019) and De Bellis (2009). Van Raan (2019) defines bibliometrics this way: “Within scientometrics, the research on scientific communication, in particular with data from publications, citations, and journals, is called bibliometrics”, with scientometrics being the “quantitative study of science”. A comprehensive discussion of the terminological questions of this branch of science can be found in chapter 1 of De Bellis (2009).

Bibliometrics as a complex of methods and tools lends itself both to the curiosity driven examination of the process of scientific research itself, by taking its communication system as a source of data, and to applied use. Applied use of bibliometrics can range from purely documentary and informative statistics to evaluative assessment with important science-political consequences, such as funding decisions.

As a largely empirically operating discipline, bibliometrics requires data of sufficient data quality in order to produce meaningful and useful knowledge. While this may seem like a truism and while the same holds true for any other data-driven discipline, data quality and its ramifications are nevertheless not routinely explicitly addressed in the substantive bibliometric literature despite a wealth of methodologically oriented studies showing severe and wide-ranging

intervention in science funding is a good idea (Kealey & Micheals, 2019).

data quality deficits of principal bibliometric data sources. Gauch (2021) goes so far as to state that “most bibliometricians are strongly inclined towards retaining a high level of data quality in the bibliographic data they use. Sometimes to an extent that may appear pedantic to the average user of these data sources” (p. 163). Perhaps, but analyses of the sensitivity of the results of studies to errors in the data are the exception in the field. Because not much is known about how much fundamental quality attributes of bibliometric data affect bibliometric indicator values, it is not possible to say whether data quality is generally good enough to make studies reliable and trustworthy or so poor as to call many results into question. Be it one way or the other, as long as this knowledge deficit exists, bibliometrics as a scientific discipline is wide open to criticism for relying on data of poor or even unknown quality, which in itself presents a vulnerability to its acceptance. In short, the overarching postulate of this dissertation is that the credibility of individual bibliometric studies, of bibliometrics as a research endeavor, and of bibliometrics-based research evaluation rests inextricably upon sufficient data quality.

The research contributions presented in the substantive chapters of this dissertation demonstrate empirically that bibliometric data quality of commonly used data sources with commonly applied methodology is often poor. So poor as to question present bibliometric practice, which is largely indifferent to appreciation of uncertainty in results caused by data quality. The contributions of the dissertation go beyond this in proposing and demonstrating methods to improve data quality and a method to propagate and express information on uncertainty of bibliometric indicators that arises from data quality. The remainder of this introductory section will provide the necessary context for the substantive studies and the motivation for studying bibliometric data quality.

1.3 Conceptions of data quality

1.3.1 Definitions

Parts of this dissertation constitute applied data quality research. That is, empirical research on the quality attributes of specific data and on the consequences of data errors for the use of this data. What, precisely, is meant by the term ‘data quality’? We proceed to discuss the concept of data quality and some dimensions of data quality that are considered of central importance in

the literature and that are also important in bibliometrics.

The international norm ISO 8000-2:2022 defines ‘data quality’ as the “degree to which a set of characteristics of data fulfills requirements” (3.8.1). ‘Data’ is defined (3.2.2) in the same source as “reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing”, while a ‘requirement’ is a “need or expectation that is stated, generally implied or obligatory” (3.2.1).

1.3.2 Demands of data quality in basic versus applied bibliometrics

An understanding of the meaning of data quality in bibliometrics can build on this basis. Bibliometric data relate to the human activity of scientific research, in practice often with an emphasis on data relating to outcomes or end products of this activity – scholarly communication. Requirements for bibliometric-scientometric data can be derived from the objectives of bibliometrics. Bibliometrics is not one indivisible category of activity. It has developed into a wide-ranging and complex sphere of activities with many facets and aspects, competing schools, conflicting motivations, and diverging orientations. Like with other scientific disciplines, some parts of this complexity can be simplified by focusing on one of several dimensions by which bibliometric research may be ordered – the basic research-applied research dimension. In bibliometrics, both investigative basic research, which aims to further our understanding of the workings of the human social activity of scientific research, coexists with applied research. This is not to say that any bibliometric study must fit somewhere on a pole or a specific point in between the poles on this spectrum. Any given study may have aspects of both basic and applied research. However, frequently a clear positioning of studies on the basic-applied continuum is possible and may be useful with respect to considerations of data quality.

In the basic, fundamental, curiosity driven area of bibliometric investigation, the goal is the attainment of relevant knowledge that is rather general and generalizable, not tied very closely to a very specific context in time, place, and culture. Here, the sought after knowledge often is in the shape of broad law-like patterns that recur again and again, statistical regularities and relationships, assessed as to their fit to some theory or demanding a theoretical explanation in the first place. For this type of bibliometrics, a certain degree of data quality deficiency may be acceptable, if it can be assured that the results are robust

enough. If any observed patterns are volatile to minor data quality deficits, this might even arouse the suspicion that they are idiosyncratic artifacts that cannot be reproduced in other contexts. Nevertheless, a fundamental level of good data quality is indispensable for generating useful basic bibliometric research results. What that required level of data quality is, may depend on the local requirements of the specific study, for instance the type of research producing units which is analyzed, where a somewhat lower data quality may be acceptable for studies of entire countries, compared to studies of individual researchers.

In contrast, when it comes to applied bibliometrics, i.e. research evaluation, assuring very high data quality is of utmost importance, as bibliometric reporting is expected to provide a reliable basis for decision-making. It is the responsibility of the bibliometric evaluation expert not only to make that assurance but also to clearly communicate the level of data quality, including its affordances and limitations, to the clients or decision-makers (Vinkler, 2010, p. 4). The values of bibliometric quantities must in this context be as accurate (and transparent) as possible to support fair decisions in high-stakes situations like funding distribution or personnel hiring, promotion, or compensation. In summary, the data quality requirements in applied bibliometrics can usually be expected to be much higher than in basic bibliometric research.

1.3.3 Dimensions of data quality

Data quality is commonly seen not as a monolithic, but as a compound concept with an internal structure. Batini & Scannapieco (2006) describe data quality as a multifaceted concept comprised of several dimensions. The most important data quality dimensions considered in the literature are accuracy, completeness, currency, and consistency. The (semantic) *accuracy*³, or correctness, of a data value is understood as the closeness between a value as it exists in an information system and a value which correctly represents the real phenomenon that the data value is a surrogate for, i.e. its true value (Batini & Scannapieco, 2006). For example, in bibliometrics, if a bibliometric database gives the publication year of an article as 2010 but from the original publisher data it can be determined with certainty that it was published in 2009, then the accuracy of that value is affected. Unwarranted duplicates of records in a database are another

³In contrast, the syntactic accuracy of a value refers to whether a specific value in an information system is a member of the admissible domain of values for a variable in that system.

instance of accuracy problems commonly encountered in bibliometric data (e.g. Valderrama-Zurián et al., 2015). These may result from errors in data collection or processing. One example are errors in data disambiguation, such as in author name disambiguation, where one real person is inaccurately represented by more than one record.

The *completeness* of data refers to missing values that should be present because the entity or attribute exists in the real world and falls under the scope of the information system. Completeness is just as crucial for bibliometric data as for any other data. Consider for example the consequences in evaluative bibliometrics if values for attributes such as author name or author affiliation are missing in the information system which do exist on the original publications. The accuracy of values can only be determined for non-missing values, the accuracy of missing values is indeterminate. This data quality concept of completeness should be clearly distinguished from the more commonly used concept of data source completeness, or coverage. The latter refers not to a data quality dimension but to a property of a data source, namely to the proportion of a specific defined literature of interest that is included as primary indexed records in that data source. Bibliometric data sources usually have accessible registers of the covered literature. There is a situation when these two concepts intersect. If a data source claims coverage in its register of literature sources but does in fact not include all items of a source in the allegedly covered period, say all articles in a journal that is supposed to be covered, then this data source has a completeness insufficiency in both meanings of the concept.

Attributes of real-world objects change over time and so must the data representing them in order to be faithful to reality. Thus, *currency* is another relevant data quality dimension. Currency means how up-to-date the data is; whether it is current or outdated. Important examples for bibliometrics are updates to publication status such as retractions or the representation of newly introduced attributes for existing records. But also just the time it takes for newly published publications to be reflected by new records in bibliometric databases is an important data quality concern for users. A value that is no longer current is necessarily inaccurate, so currency may also be seen as a sub-category of accuracy.

Consistency refers to whether data in an information system or across information systems that are to be integrated are logically consistent with each other, that is, free of contradictions or conflicts. In many cases, requirements for data

consistency can be directly integrated as logical rules in a data system must that related data must fulfill to be considered technically consistent. Consider the fictitious example of a newly founded university, active as of 2025. If we find in a bibliometric database publications with affiliation data for that university from before 2025, this may indicate a breach of consistency.

Several more data quality dimensions have been proposed in the literature, cf. Batini & Scannapieco (2006). Some are slight variations on the major consensus dimensions treated above, some are domain-specific conceptions of data quality dimensions. Thus, the foregoing is not meant to be exhaustive but selective for the purposes of this dissertation. This dissertation is mostly concerned with the quality dimension of accuracy (in the strict sense), although completeness will also play a role in one chapter: Chapter 3 deals with publication month data, which is often not available in the used data but can be approximately inferred. A method to do this, and so to overcome the incompleteness, is proposed and evaluated.

1.3.4 Other distinctions of data quality in bibliometrics

Besides these internal dimensions of the concept of data quality, several other distinctions or categorization of data quality are relevant in bibliometrics.

Bibliometric data quality studies may examine the quality of scientific publication data at different *stages*, such as the original publications, their representations in the metadata available at publisher websites, the data on those publications in citation indexes or other information systems, or the derived data that bibliometricians and scientometricians create from those sources for their own specific purposes. In this dissertation, several of these stages will play a role. It will mostly examine data as it is presented in the raw data of citation indexes. However, I will also compare this to the more basic data from publisher systems and original publications in order to empirically assess levels of data quality. Moreover, also the further processed derived data that specialized bibliometric methods produce will be studied for its quality.

A very closely related, yet logically distinct way of thinking about bibliometric data quality is to differentiate according to the *source* of errors. An error, or inaccuracy, that is present in the original publication, say for example when authors cite a reference incorrectly, may be present at later stages of data use,

but it may also be corrected in later stages. The processing system of the Science Citation Index has had capabilities to detect such erroneous references and to correct them for a long time (Garfield, 1983). But such algorithmic manipulations may themselves introduce errors. This is how the distinction between stage of data and source of error becomes clear. The source of error may be a human or software. Different error sources may act at different data stages. The specific source of error will play a central role in Chapter 4, where my co-authors and I compare different affiliation disambiguation systems which in principle use the same basic data but generate outputs of different data quality.

Data quality takes on a higher level of complexity in data source integration, when multiple information systems are combined to supply a single end user access level. Daraio et al. (2022) propose an integration-layer approach to control data quality in data source integration, focusing on the data quality dimension of completeness, whose role in integration is highlighted. Their case studies from the realm of science policy studies data illustrate that data quality control at the integration layer relieves end users of considerable data quality analysis work by providing them with harmonized data that is by design as complete as possible given the completeness of the integrated systems.

Data errors in bibliometrics can also be classified into *random errors* and *systematic errors*. When errors are randomly distributed, they affect the possible precision of calculated indicator values. If data are affected by many random errors, any quantities calculated from them will have low reliability. Data of too poor reliability may be useless. Knowing the level of data quality –and hence reliability– is therefore essential. Systematic errors have some non-random component, and they cause specific deviations of observable calculated values from the values that could be obtained if all errors were corrected. As an example from bibliometrics one might think of the core processing step of reference matching. If a reference matching algorithm is relatively unsophisticated, it might systematically match some types of cited references worse than others. Author names are a crucial component of reference strings. Think of authors with complex names, such as compound names, names with accented letters, names from different writing systems which could be transliterated in different ways into Latin characters. Such complex names, and since names are culturally embedded, names from some cultures more than others, could lead to systematically more frequent failures of a reference matching algorithm to establish

a match where there should be one. In this hypothetical example, the citation counts for entire countries might be affected by greater systematic error, that is, they would be biased and show lower citation counts than warranted. This example was hypothetical, but the problem of systematic errors in citation matching is real. Franceschini et al. (2014) were able to identify systematically different error rates in citation matching for different publishers of citing papers.

In a sense, random errors are less severe than systematic errors, precisely because they are random and are therefore equally likely to affect any observation to the same degree and with the same probability. Systematic errors on the other hand affect observations with different probability and to different degrees. However, if the class of observations that are affected by systematic errors can be identified well and if the size of the systematic error is well-understood, systematic errors can be corrected for, while this is not possible for random errors. If, for example, citations for papers by authors with complex family names are systematically undercounted because of matching errors, this can be corrected for statistically by adding a factor to each citation count of this group of papers which corresponds to the size of the known undercounting.

This dissertation is concerned with both random and systematic errors. As an example, I study the accuracy of document type assignment in Chapter 2 and find that some document types are more likely to be incorrectly categorized than others, which is a form of systematic error. The accuracy of the citation counts is studied in Chapter 5. Imprecise or faulty referencing by authors, but also imperfect reference matching by citation index producers, can lead to unmatched references. On the level of individual publications, their degree of being affected by matching failure is probably largely random since a number of independent factors bear upon the end results. However, as the citation matching algorithms themselves are (presumably) deterministic, that is, they produce the exact same results for the exact same input, citation matching errors on the level of specific cited reference *strings* are arguably systematic – if a specific reference string is not matched or not matched correctly in one citing document, it will also not be matched in all other documents in which it occurs.

1.4 Data sources

Bibliometric studies usually draw upon rather specialized information systems, bibliometric-bibliographical databases of scientific publications. Their distinc-

tive characteristic is the inclusion of publication metadata of particular relevance for science studies, first and foremost citation data, which sets them apart from similar scientific literature database systems, such as research library catalogs. The citation indexing database, as a data source of bibliometric studies, will be the location at which data quality and bibliometric indicator uncertainty are studied in this dissertation. That is not to say that the data processed, recorded, and stored in these systems –metadata on the scientific publication system– is the only relevant type of data for bibliometric studies. Other types of data, especially data collected specifically for bibliometric research, such as a survey data, are also important. However, the scope of this dissertation is limited to bibliometric-bibliographical databases, or scientific citation index databases, which is justified by their historical and present prominence as primary data sources for bibliometric activity.

Data quality deficits relevant for bibliometrics can exist in the original publication before it is ever indexed in a bibliometric database. For instance, incorrectly cited references are a typical error category at this stage. But the bibliometric data source may also itself be the place in which errors in the data are introduced, when the originally present data are not correctly reproduced. This is because the processing of original data is not an error-free process. Empirical research has shown that bibliometric databases introduce numerous errors into the data, as will be discussed later on. This means that the choice of a bibliometric data source for a study has important consequences for the level of data quality and reliability of bibliometric indicators of that study.

Large scale bibliometric research and evaluation was for decades dependent on the Citation Index products of the Institute for Scientific Information (ISI), later Thomson Reuters, currently Clarivate Analytics. ISI's founder, Eugene Garfield, created the first multidisciplinary science information retrieval tool built around the idea of searching by cited references and citing documents, the Science Citation Index (SCI) (Garfield, 1955, 1964; Wouters, 1999). Citation Indexes were in time introduced for the social sciences, arts and humanities, proceedings literature and scientific books. These have been used since their early days for bibliometric studies of the science system and for decision-making.

Data quality was taken seriously at ISI from the very beginning of the SCI's history. This effort was led by Irving Sher, Director of Quality Control and Director of Research and Development (Garfield, 2001). Sher and colleagues reported on the implementation of a procedure for correcting references cited

incorrectly in papers in Sher et al. (1966). The following quote may give an interesting impression of what data quality work meant in the 1960s:

In the preparation of the IBM cards for the Science Citation Index, all keypunched data are verified by a second operator or reader. The IBM 557 interpreter with proof feature is used to interpret the cards and edit for erroneous multipunch characters and a preliminary field edit. After all the errors detected by the verifier have been corrected, an IBM 360 computer is used to edit and identify any further errors of the type which can be checked for by algorithm. This very elaborate computer procedure confirms card sequencing, field positioning, and acceptability of the types of data allowed in the various field positions. (Sher et al., 1966, p. 134)

Among the various relevant metadata fields, it is particularly the cited reference strings, with their great variability across disciplines and journals, that are challenging to process:

Normal citation behavior in hermeneutics, the interpretation of Scriptures, includes exegesis, the detailed documentation of sources. This is also expected of lexicographers, etymologists, and encyclopedists. Similarly, law review journals provide a level of detailed documentation that might be considered pedantic in other fields. [...] On the other hand, in engineering journals, pageless citation of textbooks, handbooks, and data compilations is frequent. In chemistry, the specific pages are cited but complete pagination for the full article will be missing. In citation indexes this results in the artificial separation of citations to the same article. These individually cited pages are listed as separate items, inflating the size of the indexes (Garfield, 1996, pp. 453–454)

The last few decades have witnessed massive changes in the citation index market. Commercial competitors such as Scopus (Baas et al., 2020) and Dimensions (Herzog et al., 2020) have been established. These databases are also used regularly as sources for bibliometric studies. The vendors offer, for subscription fees, raw data access to their products and access to bibliometric analytics platforms. Data from Google Scholar, a bibliometric data source that does not

require paid subscription, has also been used for studies, with noted caveats. Completely open citation data sets have recently been made available (Peroni & Shotton, 2020; Priem et al., 2022).

Several countries and regions have set up their own citation index databases to address their local needs. There exist or have existed the Russian Citation Index (Gorin et al., 2020), the Chinese Science Citation Database (Jin & Wang, 1999), the Chinese Social Sciences Citation Index (Hua, 2001), the Arabic Citation Index (El-Ouahi, 2023), KCI Korean Journal Database, and SciELO, covering Latin American research output – all hosted on Web of Science. SciELO is actually much more than just a regional Citation Index. While its citation index component is integrated with the Web of Knowledge since 2013, SciELO was founded in 1997 and it also exists as a freely accessible stand-alone platform. This platform also encompasses an entire open access publishing system. There is further the Serbian Citation Index (Šipka, 2005, also including a journal publishing system), Citation Index of Kazakh Publications (Suleimenov et al., 2009), the Indian Citation Index (Giri & Das, 2011), the Taiwan Citation Index - Humanities and Social Sciences⁴, the Citation Database for Japanese Papers (Negishi et al., 2004), the Islamic Citation Index (Mehrad & Arastoopoor, 2012), and MyCite (Malaysian Citation Index)⁵. As of 2022, the Armenian Citation Index was under construction (Sargsyan et al., 2022). Little is known about the data quality of these regional and national citation databases. Nearly all data quality studies have used the major international commercial databases⁶. Discipline specific citation indexes also exist, such as INSPIRE-HEP and Astrophysics Data System.

1.5 Indicator use and the importance of bibliometric data

In this section a broad overview substantiated by diverse examples is given to demonstrate the motivation for research into bibliometric data quality. It is argued that bibliometric data quality should be seriously studied because bibliometric performance indicators and bibliometric data have had a profound

⁴<https://tci.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=5gvP6n/aboutdb>

⁵I cannot confidently conclude that I have not overlooked further citation indexes.

⁶However, the data quality measures of SCIndeks - Serbian Citation Index for cited reference accuracy are notable. They include, in the integrated publishing system, automated checks of correspondence of in-text citations to the reference list and a system to alert authors to references to retracted papers, cf. <https://scindeks.ceon.rs/static/about.aspx?url=qualitycontrol.htm>

transformative influence on the way science is assessed and therefore also indirectly on how science is done and which science is done. Bibliometric evaluation under conditions of insufficient data quality can thus have considerable adverse consequences. Careless use of bibliometrics may lead to flawed or overconfident decisions made unwittingly on the basis of data that, when scrutinized, did not actually allow for clear-cut decisions or would have supported different ones. In other words, particularly in an evaluative context, the use of bibliometrics in a state of ignorance of underlying data quality may be an instance of irresponsible research assessment.

1.5.1 Non-opposition of qualitative and quantitative research assessment

The appreciation, understanding and communication of the value and relevance of scientific works and of individual researchers on the basis of their works can only be done to the general acceptance of those whose research is being judged by colleagues who are highly knowledgeable about the research field and whose conduct in evaluation is so transparently objective and fair-minded to be beyond reproach. Any detectable degree of partiality, be it favorable or antagonistic, is usually regarded as a breach of the norms of science as a social endeavor with the objective of the production of certified and reliable knowledge about the world.

The in-depth, qualitative evaluation of research outputs has limitations. When the necessity arises of assessing and comparing the research outputs of larger, productive research units and organizations, such as universities, or countries, formal peer review does not scale (Abramo, 2024, p. 4). However, as bibliometrics takes over when it comes to the assessment of thousands or hundreds of thousands of research publications, neither formal peer review nor its premises are left behind completely, rather it is supplemented by products of another type of appraisal by researchers.

It is apposite to reflect here on the sometimes overlooked connections between qualitative peer assessment and quantitative publication and citation analysis. Over the past decades, a social reform movement has been gaining ascendancy which is contesting the relative roles and balance of metrics-based and qualitative research evaluation and which has found expression in a number of successive declarations, manifestos, and initiatives (Rushforth & Hammarfelt, 2023;

Sivertsen & Rushforth, 2025). Within this movement, some have argued that qualitative peer assessment is generally preferable to bibliometrics in research evaluation, a proposition which I consider unwarranted given the identity of underlying cognitive processes.

What researchers do when they read scientific literature is always also a judgment process, an internal, informal, personal assessment. Researchers are, as Polanyi (1962) put it, engaging in a constant process of ‘coordination by mutual adjustment of independent initiatives’. This assessment proceeds by shared standards of judgment. The outcome of the assessment of published research by reading researchers may be a decision now or later to cite or not to cite the work that has been read or to recommend it to colleagues, depending on its perceived usefulness for their own research, which ties citations closely to the concept of research quality (Gläser, 2024). Such decisions will be made with some of the same motivations and by the same criteria as the assessment of works in formalized peer review settings, be it manuscript peer review before publication or post-publication peer review for purposes such as recruitment, promotion, institutional or national research output assessment exercises. The two different processes have in common the valuation of the methodology and findings of research according to their scientific soundness, originality, usefulness (Aksnes et al., 2019; Gläser, 2024; Polanyi, 1962) – but also interestingness, controversy, socio-cultural relevance, ideological alignment, etc. The difference between the settings is that the referee in the formal setting is making an assessment following such criteria as mentioned but addressed to and prepared for either a small group of decisionmakers or the wider scientific community, in a more or less standardized or at least normatively circumscribed written form, and detached from their own directly relevant information needs, while the scientist as reader and citer makes the assessment along the same criteria with regard to the direct present or possible future use for their own specific research agenda and projects.

Having such a common core of assessment criteria, aligned more or less closely with idealistic scientific norms, it is understandable that qualitative peer review and quantitative bibliometric assessment methods have some intriguing parallels. Just one well-known example of such a parallel will have to suffice at this point. In peer review, great care is taken to not select any referee who might have ulterior motives. Thus, family members and partners of assessed persons are usually not eligible, or, if already members of an assessment panel, are asked

to recuse themselves in actual or merely potential cases of conflicts of interest. The parallel in bibliometric practice is the notion of excluding author-self citations (or other types of self-citation, e.g. of journals, organizations, countries) from citation counts to rule out any bias arising from differential rates of self-promotion.

Bibliometrics processes and aggregates these citations as public traces of informal, hidden, private judgments. However, and that is its main limitation, contrary to peer review, bibliometrics does not have access to and incorporate the substantive meanings that scientific works are communicating. Wouters (1999, pp. 198–204) describes this dependency in terms of two information cycles in science: the peer review cycle and the citation cycle, the latter of which latches onto the former but also influences it. Compared to qualitative peer assessment, generic bibliometric citation analysis remains an analysis of formal surface-level features⁷.

Bibliometrics is also intertwined with formal qualitative journal peer review. Ordinary publication counting re-uses the peer review and editorial decisions on manuscripts submitted to journals. This happens when bibliometric studies either select journals or other publication channels (publishing houses of monographs, conference series) based on considerations of the scientific quality standards upheld at these venues or when they rely on the inclusion and exclusion decisions about such venues which are made by the editorial source selection boards of established commercial bibliometric database vendors. Besides inclusion/exclusion selection, a refined approach may also involve the differential weighting of the publication venue in publication counting according to some criterion that may again be based on the collective judgments of the scientific community at large. By these means, fuzzy, not completely reliable, thresholds of discrimination of scientific output by approximate scientific value may be drawn whose numeric values are the results of a great many independent qualitative appraisal events.

In other words, bibliometrics, particularly evaluative bibliometrics, is the scalable analysis of the traces of myriad individual assessment decisions of researchers. In this sense, in one way the quality of bibliometrically informed valuation and decision making is totally dependent on the quality of the under-

⁷By ‘generic bibliometric citation analysis’ I mean citation *counting*. Some access to more nuanced meaning expressed in papers about cited documents is attained by the techniques of citation *context* analysis.

lying micro-valuations: to the degree that the qualitative assessments are made either more conscientiously or more arbitrarily, bibliometric methods which batten onto and aggregate data derived from these decisions gain or lose validity as indicators for impact or quality of research. In another sense, it transcends their idiosyncrasies and the multiplicity of influences of random factors affecting any particular subjective judgments precisely by its scaling and aggregating. It is only by the holistic joint analysis of many publications and citations that bibliometric indicators achieve their utility.

Bibliometric indicator scores are always based on the cognitive processes of a multitude of researchers rather than on those of only the selected few. Bibliometrics thus has intrinsically a much higher cognitive, cultural, and socio-demographic diversity than qualitative peer review exercise can achieve. It also has a different power dynamics. Namely, distributed and pluralistic, rather than the inevitably centralistic and *dirigiste* dynamics of qualitative peer review. And yet these arguably desirable characteristics only come about because bibliometrics is downstream of qualitative cognitive appraisal processes. Attempts at a Manichean juxtaposition of ‘qualitative’ peer review against ‘quantitative’ bibliometrics, of playing one out against the other, can hence be dismissed as either ignorant or misguided or fall under the suspicion of being motivated by vested interests (cf. Abramo, 2024).

1.5.2 Beginnings of bibliometrics

The practice of publication counting with the goal of identifying and studying important researchers and research nations goes back at least to the late 19th century and the *Catalogue of Scientific Papers*. The usefulness and validity of the method was immediately disputed (Csiszar, 2017).

Citation counting became accessible as a technique only with the innovation of the Science Citation Index and this innovation occasioned a momentous shift. The innovation is, as argued by Wouters (1999), the representation of science through scientific literature in a new form. This new form is created by the process of inversion of the cited references into citations. Citations connect a scientific work to the works that use it as a reference. Citations are the elementary objects of citation indexes and they are created by the construction of the citation index. Citation indexes organize citations and make them accessible, objects which apart from a citation index are not realized but would remain po-

tential without it. According to Wouters (1999), the scientist, by referencing a paper, does not give a citation. The citation is created by the citation indexing process that the scientist’s paper undergoes. References are the “raw materials” for citations (Wouters, 1999, p. 112).

Evaluative use of the Science Citation Index began soon after its first publication. Hager et al. (2024) studied the consequences of the publication of the printed SCI in the first years of its existence, the latter half of the 1960s, on U.S. researchers. By exploiting the fact that the early SCI was not published continuously and journal coverage expanded quickly in that period, they are able to make the crucial distinction between citations that were contemporaneously visible in the SCI and thus could have influenced professional researcher careers, and invisible citations, those which were retrospectively added to the SCI and cover the same period but were not available at the time. They show that it was the citations visible in the SCI volumes at that time, but not the invisible ones, that influenced careers. Scientists with more visible SCI citations were better able to secure jobs at higher ranked university departments in their discipline and less likely to drop out of science altogether compared to peers with less visible citations, even given equal numbers of publications. The same is not true for invisible citations. Hager et al. (2024) find that the introduction of the SCI had a large and almost immediate effect on the assortative matching by citation rank between researchers and departments. They also affected promotions to full professor and receiving an NSF grant. This was possible because SCI citation counts, which in the print versions were relatively easy to cumulate by first author names, revealed information on researcher quality that was not possible to obtain before⁸.

Wade (1975) notes several applications of the printed SCI in assessment and research in 1975, including promotion and tenure decisions in universities. According to Garfield (1979, p. 62), “the *SCI* data base is being used to do such things as evaluate the research role of individual journals, scientists, organizations, and communities”. In the U.S., the use of bibliometric indicators for departmental decision-making was well established by the 1990s. Hargens & Schuman (1990) report survey data according to which 35 % of biochemist and

⁸This account of the immediate evaluative use of the SCI is in contrast to Wouters (1999), p.4: “Nevertheless, the scientific community was not enthusiastic. Many researchers did not even use the SCI for its stated purpose as a bibliographic tool – to find relevant publications in the exponentially growing mountain of scientific literature. Neither did many researchers use it to keep abreast of their citation status, a measure without clear meaning to many scientists. The prevailing reaction was hostile or indifferent.”

60 % of sociologist respondents' departments had "made use of citation counts in making decisions about hiring, promotion or salaries." Experimental bibliometric monitoring of research groups began in the 1980s at Leiden University (Moed & van Raan, 1988). The importance of bibliometrics has only increased since then.

1.5.3 The present role of bibliometrics

Currently, metrics whose application on the level of individual researchers are highly controversial in the bibliometric research community, such as the Journal Impact Factor or the h-index, are widely used for academic personnel decision making (Hammarfelt & Rushforth, 2017; McKiernan et al., 2019; Rice et al., 2020). For example, a survey of medical researchers found that the number of publications, first authorships, and the Journal Impact Factor, but not the number of co-authors of publications, were relevant criteria in their performance reviews (Walker et al., 2010). Citation metrics are frequently invoked as relevant evaluation criteria in documents relating to staff review, promotion and tenure in North American universities, albeit sometimes critically (Alperin et al., 2019). A recent Norwegian study has found that for professorial recruitment bibliometrics are used mostly as a first-stage screening tool to reduce the number of candidates for in-depth qualitative review (Reymert, 2021).

While there is support for some application of indicators in personnel decision making among scientists, preference for metrics varies along several dimensions: those researchers who are rated more highly in research quality show stronger metrics preference than those rated relatively poorly; women prefer metrics less than men; STEM researchers are endorsing indicator use more than social science and humanities scholars (Cruz-Castro & Sanz-Menendez, 2021). Attitudes to assessment are then hardly free of motivated reasoning or self-serving biases. As a case in point, the participation in a boycott of a public ranking of German business scholars was shown to be related to how much individual researchers had to gain or lose in such an ongoing change of evaluation culture towards publicly highly visible metrics (Berleemann & Haucap, 2015).

Besides the aforementioned individual researcher assessments, bibliometric indicators are used for all kinds of applications, from use in grant application by review panel members (Gunasekar et al., 2017; Lewison et al., 1999), to national-scale comprehensive research evaluation systems (Hicks, 2012). One

example for the latter is Italy, where the introduction of citation indicators is thought to have led to considerable gaming of bibliometric indicators (Baccini et al., 2019). Performance-based funding systems all over the world rely on bibliometrics. For example, Hammarfelt et al. (2016) found that “almost all major HEIs [higher education institutions] in Sweden use bibliometric measures for resource allocation”. Most German medical faculties use bibliometric measures for funding allocation (Aman & van den Besselaar, 2024). Performance-based funding in Australia on the national level and at universities uses bibliometric data and methods (Woelert & McKenzie, 2018).

There are several national and international research monitoring report series that make extensive use of bibliometrics. Prominent examples are the the U.S. National Science Board’s *Science Indicators*, which started in 1973 (National Science Board, 1973), various reports by the EU, the latest incarnation being *Science, research and innovation performance of the EU, 2024* (European Commission and Directorate-General for Research and Innovation, 2024), and the *Research Fronts* series by the Chinese Academies of Science and Clarivate. Germany uses bibliometric reports for progress monitoring of its major policy instrument for financing extra-university research, the Pact for Research and Innovation⁹.

Perhaps the most publicized use of bibliometric indicators is in global university rankings. All major rankings have been using progressively more sophisticated bibliometric measurement as one component among several others.

This by now ubiquitous metrics-saturated environment has lead to a certain degree of socialization such that use of bibliometric indicators has become a part of scholars’ everyday work, as shown for the social sciences in Australia and Sweden by Haddow & Hammarfelt (2019) or for biomedicine in the Netherlands (Rushforth & de Rijcke, 2015). Haddow & Hammarfelt (2019) characterized this usage as “pragmatism and compliance” (p. 23). Nevertheless, detailed knowledge about the technicalities of indicators is limited among researchers (Kamrani et al., 2021).

Despite fundamental objections, despite frequent inappropriate use, despite limitations pointed out by bibliometricians, despite a surge of both top-down and bottom-up initiatives to rely more on societal impact and non-metric assess-

⁹<https://www.gwk-bonn.de/themen/foerderung-von-ausseruniversitaeren-wissenschaftseinrichtungen/pakt-fuer-forschung-und-innovation>

ments of research – bibliometric indicators have become ever more widespread in application and it looks like they are not going to go away any time soon. While this encroachment and proliferation is frequently objected to, lamented, resisted, and attacked, the data quality of the data basis of metrics-based assessments is surprisingly rarely raised as an argument in these discussions. However, the professional bibliometric research community has in fact devoted some of its attention to this matter. The next section will sketch a selective overview of major topics of this research literature.

1.6 Bibliometric measurement uncertainty as a research subject

Validity and reliability are central issues in the assessment of the usefulness of bibliometric measures for evaluation of scientific performance. (Nederhof, 1988, p. 194)

Validation studies may help to bridge the already existing gap between bibliometric research and technology. Beyond technical validation studies which are designed for detecting systematic and estimating random errors, methodological validation studies may ensure that we really measure what we are intending and assuming to measure. (Glänzel, 1996, p. 171)

population parameters can still be inaccurate due to measurement error, an issue seldom discussed in relation to citation indicators (Schneider, 2013, p. 58)

This section will review research topics in bibliometrics related to uncertainty, data quality, and measurement. It will serve to contextualize the original research of the dissertation. The substantive chapters of this dissertation are all concerned with the role of uncertainty in bibliometric indicators. While uncertainty contingent on data quality is the overarching topic, parts of the dissertation are also concerned with the related topic of measurement validity. One important distinction has to be drawn at this point. Science itself is full of uncertainty. Scientific outcomes of experiments, of research projects, of whole research programs, are uncertain. That is why the research has to be done. For

Leydesdorff (2001), science is essentially the systematic reduction of mankind’s uncertainty about the world. This kind of epistemic uncertainty I must set aside as being outside of the scope of this dissertation – it will only be concerned with measurement uncertainty in bibliometrics.

Known uncertainties are preferable to ignorance about uncertainty. Ignorance about uncertainty can lead to unwarranted belief in the trustworthiness and accuracy of findings, which can lead to misguided decisions. Bibliometricians sometimes call for explicit assessment of uncertainty in results. For instance, in a proposed method for evaluating global university rankings, Gadd et al. (2021), under the rubric of “rigour”, postulate the criterion of *honesty about uncertainty* that rankings should meet: “The types of uncertainty inherent in the methodologies used, and of the data being presented should be described, and where possible, clearly indicated using error bars, confidence intervals or other techniques, without giving a false sense of precision”. Another instance was in the Leiden Manifesto (Hicks et al., 2015), which recommended to “Avoid misplaced concreteness and false precision [...] If uncertainty and error can be quantified, for instance using error bars, this information should accompany published indicator values.” The present dissertation contributes to the development of the understanding, appreciation, quantification, and expression of uncertainty in bibliometrics.

1.6.1 Elements of uncertainty

Uncertainty in scientific results has several distinct components. The first one to consider is sampling uncertainty. Data that is randomly sampled from a population is subject to ordinary sampling error, leading to uncertainty in population parameter estimates. Sampling uncertainty is the subject of the theory of probability and the arsenal of inferential statistics. Bibliometric data often do not constitute a random sample from a larger population. Rather, the data of the population is available or the available data is in no sense a randomly sampled part of the population. In these cases, expressions of statistical sampling uncertainty are unnecessary and inapplicable (cf. Schneider, 2013, for the context of statistical significance testing)¹⁰. When bibliometric data, however,

¹⁰It has been argued that the citation process is probabilistic, that it has stochastic elements (e.g. Nederhof, 1988, sec. 2.2.3), and that therefore the citation count of a publication has some statistical uncertainty, which can possibly be modeled mathematically (Dieks & Chang, 1976). This line of reasoning has not been worked out further nor found application in later

are randomly sampled, the appropriate application of inferential statistics to quantify sampling uncertainty is called for. As an example, in Chapter 2 of this dissertation, I randomly sampled bibliographic items stratified by document type to study document type category assignment error rates. Since these are true random samples of a clearly defined larger population, it was possible and meaningful to calculate probabilistic bounds for the population error rates from the observed sample error rates (confidence intervals). The major source of uncertainty in bibliometric results is usually not statistical sampling. Rather, most of uncertainty in bibliometrics arises from issues of measurement reliability (including data quality) and measurement validity.

We now move to a discussion of the relationships between core concepts of this section. The concept of data quality was already treated in section 1.3. The other concepts are measurement reliability, measurement validity, and measurement uncertainty. As these concepts can be understood to form a chain of dependencies in this specific order, the discussion of their relationships proceeds by that sequence.

In bibliometric research, measurements are made from pre-existing or newly collected data about the scholarly communication system. The quality of this data, primarily its accuracy, may be more or less defective, as will be shown below. The quality of the used data is one factor that affects the *reliability* of measurement. Reliability (of measurement) refers to how accurately and consistently some quantity or attribute can be measured, in other words “the absence of measurement errors” (Alwin, 2005). One useful consideration which is very applicable in bibliometrics is the idea that in measurement there exist both a true score (or value) of a variable, which may only exist theoretically, and that value of the variable which we can observe through measurement, the observed score. The difference between observed and true score is measurement error. A good example from bibliometrics are citation counts of a specific citation database at a specific time. Suppose the measured, or observed, value of a publication’s citation count in a database is 10. We could establish also the true value of that variable with some effort, which is what I did for a sample in Chapter 5. This can be done by manually or semi-automatically checking all similar cited reference strings and counting, independently of the database’s own reference matching procedure, those that refer to the publication in question. Suppose one did this and counted 11 such reference strings. Then the true score is 11

research.

and the measurement error, caused by imperfect data quality, is -1 . Other ways of establishing the reliability of citation counts may be relevant and have been considered. Suppose one is more interested in the reliability of citation counts independent of one specific database. One could then obtain citation counts for the same publications from the same corpus of indexed (citing) literature in different databases and compare them (Franceschini et al., 2013). An issue of reliability of citation counts is that short term citation counts, as collected a few years after publication, correlate more or less with long-term citation counts. Long-term citation counts are arguably the theoretically more appropriate way to measure scientific impact by citations, as the whole citation lifetime of publications is considered rather than just the first few years of citation history. If this reasoning is accepted and it is nevertheless necessary to count citations relatively shortly after publication to have somewhat current data, the reliability of early citation counts for lifetime citation counts becomes an important issue to study (Abramo et al., 2012; J. Wang, 2013).

While it may perhaps appear so in the bibliometric context, data quality and reliability are not the same thing: data quality is one causal factor affecting reliability in bibliometric measurement. Another might be, in the context of the citation count example, the comprehensiveness and quality of the reference matching algorithm.

Another important consideration about reliability to keep in mind is that reliability is only concerned with measured and theoretically measurable values but not with theoretical concepts or constructs. Thus, in bibliometrics, ‘citation count’ is a measurable quantity of a publication record in a citation database. But citation counts are not the same thing as the concept of ‘scientific impact’.

This brings us to *validity*. Validity, in the sense of ‘measurement validity’, refers to how well a specific theoretical concept or construct, which one would like to measure by some measurable quantities, is reflected by these measurements. It is often defined as the “extent to which an indicator of some abstract concept measures what it purports to measure” (Carmines & Woods, 2005). Another way to express the same meaning is to state that validity is the extent to which a measurement instrument’s scores reflect true variation in the construct it is supposed to measure (Whitley & Kite, 2012). Ad hoc measures, those put forward to measure a construct for a study without having their validity demonstrated, “are virtually useless for the assessment of hypothetical constructs” (Whitley & Kite, 2012, p. 174). Science studies investigates many theoretical constructs,

for example scientific collaboration, research quality, interdisciplinarity, prestige, reputation, and knowledge transfer. These are then attempted to be measured by some empirical methods such as surveys or bibliometric methods. For some constructs, validity of measurement instruments can be considered well-established (e.g., for citation counts as partial indicators of research quality, Aksnes et al., 2019, 2023; Thelwall et al., 2023), for others, a validity deficit can be diagnosed (Donner & Blümel, 2024; Q. Wang & Schneider, 2020). This sense of the term validity is to be distinguished from that of the ‘conceptual validity’ of a concept itself, that is, how meaningful a concept is as a description of reality (Locke, 2012).

Reliability and validity are logically independent concepts, but in practical terms, the validity of some measurement of a concept depends on the reliabilities of the variables that are invoked to measure it. Poor reliability of indicators means that a construct, trait, phenomenon, etc., cannot be accurately measured by a measurement process. This is why, especially in bibliometrics, considerations of data quality fall under the rubric of measurement reliability. Consider that a specific kind of data is theoretically a highly relevant indicator for a phenomenon. If it turns out that the data that was collected is affected by many errors, the measurement cannot be reliable. If the indicator cannot be reliably measured, the measurement of the concept by this indicator cannot have high validity.

Systematic measurement error can impede the measurement validity of a measured indicator for a concept. Systematic error in measurement, viz. bias, constitutes non-random variability in the measured values for a concept that is unrelated to the meaning of the concepts. One case of measurement validity in bibliometrics being lower than it could be is treated in Chapter 3: *Effect of publication month on citation impact*, so I will use this case as an example here. If the scientific impact of publications is measured by citation counts as a proxy variable, or indicator, then systematic variation in the indicator unrelated to the concept of impact should be eliminated as much as possible. Doing so is standard practice in professional bibliometrics and is usually referred to as normalization. It is not always made explicit that the reason for normalization is to remove variability systematically unrelated to a specific measured concept. It is more common to find the justification that one would like to make raw citation counts more comparable to each other, which is a more non-technical expression of the same concern. Citation counts are routinely normalized for several char-

acteristics which are argued to distort the relationship between measured values and the concept of scientific impact. For instance, if publications published at different times have to be compared, it is arguably necessary to normalize citation counts for the effects of longer or shorter duration of being published and available for citation because the citation counts of publications published earlier will be systematically higher. A difference which is unrelated to different levels in scientific impact. In Chapter 3, this argument is developed further because the standard methodology for citation count normalization only normalizes for publication year. In that chapter it is argued and demonstrated that a citation count bias also exists in favor of publications which were published earlier over those published later in the same year. It is then demonstrated that this systematic error can be eliminated by using month-precision citation windows, as opposed to the conventional year-precision citation windows. Month precision citation counts thus have higher validity for the concept of scientific impact because they remove bias in the measured variable.

Both reliability and validity are causes of uncertainty about the results of measuring a concept. Measurement uncertainty refers to how uncertain (imprecisely known) the value of an obtained measurement is, given all available information about the measurement process, including random and systematic error. The *Guide to the expression of uncertainty in measurement*, an authoritative document for measurement uncertainty in the physical sciences, defines the concept as follows: “parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand” (JCGM, 2008, p. 2), where “measurand” refers to the quantity that is to be measured. The *Guide* notes that this definition is compatible with uncertainty as a “measure of the possible error in the estimated value of the measurand as provided by the result of a measurement; an estimate characterizing the range of values within which the true value of a measurand lies” (p. 3). Measurement uncertainty is hence a quantified expression on how precisely a measurement reflects the underlying but unobservable magnitude of a concept taking into account all pertinent knowledge about all sources of uncertainty in the process of measurement.

1.6.2 A panorama of sources of uncertainty in bibliometrics

We can now continue with a discussion of the variety of sources of bibliometric measurement uncertainty that have been reported in the literature. The following discussion does not attempt to be exhaustive or representative but illustrative. The point is to demonstrate the wide range of primarily data quality issues that cause measurement uncertainty.

1.6.2.1 Accuracy, Completeness, Duplicates It is probably safe to say that all bibliographic data are affected more or less frequently by inaccuracies. Author names can appear incorrectly on publications' author lists due to mistakes in proofing by publishers (Teixeira da Silva, 2024). Demetrescu et al. (2018) compared author name form in four major bibliographic data sources with canonical variants for a large number of mainly Italian names. They find error rates (incompatible name strings) between 0.25 and 3 %. Not only authors' names, but also the information on corresponding authorship has been found to be affected by substantial inaccuracy in bibliometric databases (Chinchilla-Rodríguez et al., 2024).

DOIs in bibliometric databases are affected by errors of different types. For example, one DOI can be assigned incorrectly to more than one item or an item can be indexed with the wrong DOI (Franceschini et al., 2015a; Krauskopf & Salgado, 2023; Liu et al., 2021; Xu et al., 2019; Zhu et al., 2019). Franceschini et al. (2016) discuss several other identified data quality problems in Scopus, among them spelling mistakes in article titles, the overlooking of citation links to papers cited while available online without issue and page information, duplicate records for the same publications, cited references in the database record that are not present in the original paper, and cited references missing most of their information. Inaccurate funding data has also been reported for the major databases (Álvarez-Bornstein et al., 2017; Grassano et al., 2017; Liu, 2020). It has been found that WoS has not indexed author affiliation data that is present in the original publications (Liu et al., 2018). Commercial bibliographic-bibliometric databases have been found unknowingly to index content by fraudulent, hijacked journals (Abalkina, 2024).

One chapter in this dissertation, with the title *Document type assignment accuracy in the journal citation index data of Web of Science*, is an investigation into the prevalence of another kind of incorrect or inaccurate values. There I

study the extent to which the document type categories assigned to a sample of publications during indexing in citation databases differ from an independent intellectual assignment of document types using the definitions and criteria of the citation database. Another earlier study in which the correctness of document types played a role is that of Heneberg (2013) on seemingly uncited papers of Nobel Prize and Fields Medal laureates. Heneberg found that among the records in WoS of these laureates which were both uncited and categorized as articles or reviews, 26 % were incorrectly categorized and were actually of different document types.

Inaccurate citation counts and errors in cited references lists are a major topic of the bibliometric data quality literature (e.g., Cioffi et al., 2022; Franceschini et al., 2013; Liang et al., 2014; Moed & Vriens, 1989; Olensky, 2015; Olensky et al., 2016; Tüür-Fröhlich, 2016; Xu et al., 2019). The accuracy of WoS citation counts is also a component of the Chapter *Data inaccuracy quantification and uncertainty propagation for bibliometric indicators*. This chapter also gives a thorough overview of the literature on this topic, so it will not be preempted here.

Occasionally, inaccurate citation counts come under the spotlight of attention, such as when *Nature* staff noticed an incorrectly low Web of Science citation count of a landmark human genome paper published in their journal and wrote the following in an opinion piece:

When puzzled Nature staff examined the ISI's data, they found that citations to the paper had been grossly undercounted. The same, it emerged, applied to several other prominent papers authored by a consortium, rather than by a list of individuals. To its credit, the ISI reacted promptly, amending its website and commissioning a journalist to write an account of the episode for Science Watch. The paper by the International Human Genome Sequencing Consortium now sits at the top of the hot papers list. (Anonymous, 2002)

It is worth noting in this context that papers with group authors are particularly susceptible to indexing and citation errors (Dickersin et al., 2002).

The above mentioned study, Heneberg (2013), identified further causes of apparently uncited publications of Fields and Nobel laureates besides mis-classified document types. Of the 100 ‘uncited’ articles and reviews (according to WoS),

Heneberg was able to find citations by manual cited reference search for 46 of them, which were simply not linked by WoS due to errors present in the original citing document or introduced by WoS indexing¹¹.

When authors cite references with incorrect information, matching the cited reference information to the correct intended target becomes more difficult. The unique study of Waytowich et al. (2006) showed that the accuracy of cited reference data is related to individual differences in psychological traits. They studied PhD candidates' referencing errors and deviations from a proscribed citation style norm in dissertation proposals and these error scores' relationships to a measure of the construct of 'perfectionism', to expectations of study success, and to demographic variables. The two citation list quality scales were found to be correlated. Students with lower course success expectations made more citation mistakes and students with more completed courses showed lower adherence to citation style rules. Analyzing citation quality and perfectionism scales showed that "students who reported the lowest levels of self-oriented perfectionism and socially prescribed perfectionism tended to commit the most citation errors and constructed reference lists that departed the furthest from APA (2001) stipulations" (Waytowich et al., 2006, p. 203). While this study must be interpreted with some caution due to small sample size, it would, if confirmed by more evidence, point to the hypothesis that at the level of researchers committing referencing errors, stable processes are at work that are connected to the individual dispositions of persons.

In a fine piece of investigative bibliometric detective work, Besançon et al. (2024) have recently documented a novel manipulative phenomenon which they have dubbed 'sneaked references'. Sneaked references are the intentionally disguised counterpart to missing references. They are references which are not present in the original publication but inserted into their metadata at the registration of this metadata at Crossref and from there on propagated to information systems. The patterns uncovered by Besançon et al. (2024) indicate intentional exploitation of this system for nefarious purposes. From this example it is apparent that true data quality issues, here the inaccuracy of values, need not be caused by some error-prone processing step or negligence. They can be caused by systematic abuse of infrastructure. In this case they lead to increases in citation

¹¹This finding, incidentally, points to the relative fragility of indicators of uncitedness. Another reason to be skeptical about uncitedness indicators, also deriving from their dichotomous nature, is that is probably the easiest citation indicator type to influence by self-citation.

counts of cited papers, the cited references of which are hidden.

1.6.2.2 Entity disambiguation, reconciliation, matching, deduplication Various entities that are treated as actors in the socio-cognitive system of scientific research are not readily identifiable by looking only at their names. Entities such as authors, research performing and funding organizations (and their subdivisions), countries (and their subdivisions), journals and other publication venues, and publishers all appear in the written documentation of scientific research with various different surface forms of their names. This includes, of course, alternative names, changed names, abbreviations, translations and transliterations, spelling mistakes, and probably more. Different persons and organizations may happen to share a name or name form. All these ambiguities, only some of which are data errors, have to be resolved by linking the various surface forms of an entity to its identity or a preferred form, a process that information science describes by the somewhat ominous term of authority control and that computer science calls (entity) disambiguation or reconciliation. Entity disambiguation systems, usually algorithms, are themselves a source of inaccuracy since none of them work with perfect precision. The performance of such systems is conventionally judged by their level of accuracy.

The disambiguation of the names of all of these mentioned entities has been the subject of bibliometric research (Leydesdorff et al., 2016; Prancutè, 2021). In particular, the disambiguation of author names and of organization names have attracted a lot of attention, also from computer science, because of their obvious practical importance.

To give just one example, Sinatra et al. (2016) studied the timing of researchers' most impactful publication in their careers and the influence of individual level factors on citation impact. For author name disambiguation, they iteratively merge authors in an American Physical Society data set based on simple criteria, one of which is identical family name, meaning that the records for people who have changed names will never be unified. They explicitly report false positive and false negative rates of intellectually checked samples. The easy availability of this data set has prompted a fair amount of re-use. Such heuristic author name disambiguation approaches are relatively scalable while resulting in reasonably good disambiguation quality. Similar approaches are, for instance, D'Angelo et al. (2011), Caron & van Eck (2014), D'Angelo & van Eck (2020). The results of studies with disambiguated author names are systematically different to the

otherwise identical studies using unprocessed names (Erman & Todorovski, 2015; Schulz, 2016). In other words, bibliometric research which studies authors as its object and uses undisambiguated name data will be systematically biased against some authors or groups of authors and their results cannot be accepted at face value.

The names of research performing organizations, too, have to be unified if one wants to carry out meaningful research in which the objects of study are organizations. The question of disambiguation of the identity of research organizations from address-affiliation data is taken up in Chapter 4, titled *Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems*. The study compares the disambiguation results of different systems on the same bibliometric data from a perspective of applied bibliometrics. This chapter also provides an overview of related work on institutional disambiguation for science studies. A related study is that of Orduña-Malea et al. (2017), who provide an empirical study and assessment of the quality of Google Scholar’s affiliation feature – an example for the relatively rare studies of Google Scholar’s data quality. They identify several problems in the unification processes that lead to, for example, universities with similar names being conflated under one affiliation and universities being subsumed under the wrong normalized name.

Bibliometric databases or analytical data sets may comprise records from more than one external data source. If the coverage of the sources overlaps, there will be duplicate records in the derived data that need to be identified and deduplicated, that is, all but one record removed or all records consolidated into one. Valderrama-Zurián et al. (2015) studied duplicate records in Scopus. They were able to attribute such duplicates to mis-mapping of records to more than one journal, journal title changes and spelling variants, and journal name variants – these latter resulting from indexing records from multiple data sources with insufficient deduplication. Liu et al. (2021) studied indexation errors in WoS and Scopus of one particular megajournal and identified record duplicates as one type of error.

More generally, whenever data sources are combined to satisfy information needs, one faces problems of data integration and interoperability. One comprehensive approach to deal with such issues of data governance and management that has also been used for implementations geared towards evaluative and monitoring purposes for supporting science and innovation policy is ontology-based

data integration. This is a top-down approach in which the domain of knowledge is formally modeled in a machine-readable way and mappings from the ontology to concrete data sources are created. Crucial advantages of this approach, according to Lenzerini & Daraio (2019), are the ability to logically reason over the domain concept structure, the mapping, and the data. Besides that, this approach has provisions for data consistency and data quality analysis.

1.6.2.3 Completeness Data that is present in the original publications is sometimes found to be missing in the indexed records of these publications in bibliometric databases. Jacsó (2009) provided an early comparison of WoS and Scopus specifically pertaining to missing values in essential data fields. He finds a considerable missing rate for document type in Scopus, noting that “This is important to know because using the document type to filter the search results could exclude millions of wanted documents that do not have any document type information. The nine per cent omission rate [of Scopus] is too high for scientometric purposes”. The other important difference between the two data sources was the much higher rate of missing country information in Scopus, although the reasons for this could not be determined. The ramifications are nevertheless obvious for bibliometric studies using this data source: “Irrespective of the reasons for the very high rate of omission of country names or codes, it should be realised and prominently mentioned in any scientometric country reports. I have never seen this mentioned in published papers, nor in the manuscripts that I have peer reviewed.” Liu et al. (2021) reported dozens of missing records for major bibliometric databases for a single large journal.

For bibliometric studies it is often important to have not only the author names or identities and the organization names and identities which are responsible for publications but also the information about which authors are associated with which organization and country – the author-affiliation links. This data can be useful, for instance, to estimate the degree of contribution of co-publishing organizations in jointly produced work. Web of Science has indexed these author-affiliation links only from 2008 on (Birkle et al., 2020). This means that studies using WoS data that need this information cannot calculate correct values for the period before 2008.

Missing references are cited references present in the original publication but erroneously not included in the information system or citation database. They have been repeatedly observed (Buchanan, 2006; García-Pérez, 2015; Rodrigues

et al., 2023; Van Eck & Waltman, 2017).

Curiously, major commercial databases do not index some information that is useful both for information search and bibliometrics which are routinely available in published papers, such as the editorial handling dates (received, reviewed, revised, accepted dates) and the classification codes of various disciplinary classification systems. But it cannot be said that this constitutes missing information since there was never an intention to include these data.

1.6.2.4 Methodological robustness and sensitivity analysis Robustness tests and sensitivity analysis are tools to ensure that the results of a study are not artifactual and vanish when minor methodological variations are applied. They concern not only individual values of variables but the whole pattern of results of studies. Robustness checks consist of more or less systematic tests of whether findings can be reproduced under alternative operationalizations that should not overturn the findings qualitatively, such as using a different but equally applicable data source, a different population of study, alternative measurement instruments, alternative statistical techniques, etc. Sensitivity analysis concerns more specifically the robustness of results to systematic variations in the values of free parameters of algorithms that researchers have to set. Findings that remain invariant or qualitatively unchanged under such changed conditions are evidently robust to alternative specifications of a research methodology and can be said to be relatively more epistemically reliable compared to findings that do not survive such tests. When the robustness checks concern specific values and their robustness under changed conditions within a given methodology, such as variations in parameter values of algorithms or the criteria for selection of objects of observations, there will be ranges of result values for result parameters of interest. These ranges can be interpreted as the uncertainty in result values that can be attributed to minor variations in methodology (cf. Stahlschmidt & Schmidt, 2019).

The Leiden Manifesto (Hicks et al., 2015) explicitly suggested such checks: “Science and technology indicators are prone to conceptual ambiguity and uncertainty and require strong assumptions that are not universally accepted. The meaning of citation counts, for example, has long been debated. Thus, best practice uses multiple indicators to provide a more robust and pluralistic picture.” Especially when there is insufficient knowledge about the comparative validity of alternative analytical choices, one way of addressing this is sensitivity analy-

sis. However, this should not be confused with the uncertainty in results that can be derived from known validity coefficients of specific applied instruments, indicators and other methods.

A paradigmatic case is the debate on the alleged relative decline of British science in the 1970s-80s. Irvine et al. (1985) and B. R. Martin et al. (1987) opened the debate, finding a decline in the percentage of British science of the world total according to figures computed from CHI's database compiled from SCI data. The CHI databases used fixed journal sets and the methodology involved fractional counting of papers at the level of countries. The central finding, a decline in British world percentage of paper production, was challenged by Leydesdorff (1988), who reproduced the original findings but contrasted them with results showing an increase in Britain's world contribution when using the dynamic journal set of an online version of the SCI and whole counting. Kealey (1991) also criticized the journal set albeit without presenting evidence, but tried to change the debate completely by postulating that per capita numbers of papers and contribution to the couple of hundreds of most cited papers per year are what matters. These indicators presented a much more favorable view of the status of British science. Braun et al. (1989a) went some steps further and did an actual sensitivity analysis by processing the same data with different justifiable methodological choices, including journal set specification and counting method, but also others. They found "the annual mean relative change of the 28 indicators varied between -0.88% and $+4.30\%$ with an average of $+0.93\%$. Since the standard deviations within each time series were in the range 1% to 3%, no statistically significant trend could be attributed to any of the indicators." The substantive and methodological debate by no means ended there, it was extensive and detailed and continued for years (Leydesdorff, 2017; B. Martin, 1991, 1994; B. R. Martin, 2017). As this first example illustrates, methodological choices can and do have important ramifications for bibliometric findings and great care must be taken that results are robust to minor changes in methods. Even in the absence of empirical validity knowledge about methods, trying out different methods and parameters can tell us if these choices make a difference or not. This explanation makes it clear that robustness/sensitivity analysis does not neatly fit into the conceptual sequence (data quality $\rightarrow$ reliability $\rightarrow$ validity $\rightarrow$ uncertainty) of this dissertation. One may conceptualize it as a distinct side chain that itself on one side ends in uncertainty: (methodological degrees of freedom $\rightarrow$ robustness/sensitivity analysis $\rightarrow$ uncertainty $\leftarrow$...). As robustness

and sensitivity analysis are important tools to assess the overall uncertainty of results I have discussed them here. They are, however, not treated further in the substantial chapters of original research that follow.

As a more recent example of a dedicated systematic sensitivity analysis, Stahlschmidt & Stephen (2022) investigated how the different scopes of content indexation in Web of Science, Scopus, and Dimensions affect citation impact analysis in bibliometrics. By calculating normalized citation scores for publications with German affiliations indexed in the three sources, the study finds that it does make quite a difference which citation data source is chosen for the measured citation impact performance of German science system sectors. Another example for a dedicated study of the possible factors influencing results in bibliometrics is that of Struck (2025), which studied a multiplicity of causes that affect the results of citation network-based clustering of publications used for detecting and tracking topics. Some of the identified dependencies are the data source, software configuration, data model, and clustering algorithm. Another clear validity deficit is diagnosed in this area.

A crude but common method of checking for methodological robustness with regard to the influence of different methods of accounting for co-authorship in publication counting, i.e. credit allocation methods, is to run the same analysis but using different such methods and assessing if the findings remain stable. Porter (1977) (p. 260) was an early advocate of sensitivity analysis in science studies in the context of methodology of allocating credit to co-authors of multi-author papers. One recent example is Sinatra et al. (2016), who used equal fractional counting, an author order-sensitive allocation method, and a collective perception method of attributing publication credit to co-authors. In addition, they also repeat the main analysis only for authors whose most-cited papers are solo-authored, in order to rule out confounding by co-authorship in general. The robustness checks of Sinatra et al. (2016) go much beyond counting methods. They test and ensure that their main findings reproduce under many different conditions, including removing difficult to disambiguate names entirely, using different base data sets, different choices of selection of observed authors in these data sets, rescaling key citation metrics, and excluding review papers. Another study with comprehensive robustness checks is that by Hager et al. (2024), which was mentioned before in the context of the early adoption of the Science Citation Index for faculty review and hiring. The application of robustness checks in scientometric research has been raised to a new

higher standard by the introduction of multiverse analysis, a rigorous systematic approach to methodological robustness in empirical research, by Leibel & Bornmann (2026).

1.7 Organization of the dissertation

What follows are the four chapters of original research that together comprise the research content of this dissertation. The general guiding research question of this research project is:

How does data quality affect the reliability of bibliometric indicators?

The chapters are structured and organized as follows.

In Chapter 2, I study the prevalence of inaccuracies in the indexed document types of publications in citation databases and whether these mis-assignments have a measurably distorting influence on the calculation of normalized citation impact indicators. The accuracy of document type classification in scholarly information systems has never been studied before and the same holds for consequences of inaccurate document type data on citation indicators. For this study, a random sample of records was drawn from WoS, which was stratified according to WoS document type. This means that ordinary statistical inference is applicable and is used in the form of population parameter estimates with confidence intervals. Document type accuracy is assessed by manually re-classifying this sample's document types according to the database's definitions and then comparing the independent manual document types assignments to those of the database. Incorrect document type classifications will lead to the inadvertent exclusion or inclusion of unwanted publications with wrongly assigned categories when only publications of certain types are selected for a study, which is a very common step in bibliometric research and assessment (Jacsó, 2009; Waltman, 2016, p. 372). I further study if the document type classes assigned to the same papers by the two databases Web of Science and Scopus are consistent. Moreover, I also study in which way and to which magnitude inaccurate document type classification exerts an influence on normalized citation scores, which are a core element of advanced citation analyses. As normalization is generally

carried out with respect to the variables publication year, scientific discipline, and document type, some influence of document type errors is expected.

In Chapter 3, another issue related to citation count normalization is studied, namely the dependency of citation counts on publication month. The conventional method of normalization for time in citation analysis is to compare only the citation counts of papers published in the same year. The reason is that papers published in earlier years than others have had more time to accrue citations. But this problem also exists for papers published in the same year: there are large differences in average citation counts across publication months to the advantage of earlier over later months (as pointed out before, e.g., Gai et al. (2015)). In this chapter I investigate this effect with a large-scale systematic study. This includes the existence of this effect (or bias), its size, and a test of a method to remove it. As citation month data is not always available in citation index databases I also test and evaluate a method to infer this information from other publication data. This chapter is therefore concerned with the reliability and validity of citation counts and normalized citation scores as indicators of scientific impact. If the factor publication month has a systematic distorting effect on citation counts which is unrelated to the construct that citation counts are used as a proxy for, then this bias should be removed to improve the measurement validity of citation counts. Finally, I assess if the publication month effect matters in a research evaluation scenario. To this end, I test in the chapter if aggregated citation scores of German research organizations are measurably affected by publication month bias.

In Chapter 4, the issue of institution affiliation disambiguation is investigated. In this co-authored study, we first empirically establish the very high accuracy of a reference data set of records on publications from Germany. We proceed to use this data as a gold standard to compare the quality of the affiliation disambiguation systems of both Web of Science and Scopus, plus the results of using institution name searches. We analyze the performance of the affiliation disambiguation systems in retrieving the publication sets of research performing institutions, as far as covered in the respective databases, as completely and accurately as possible. This chapter also includes an analysis of the real-world effects of the possibly insufficient data quality of the disambiguation systems on a broad set of standard bibliometric indicators from the three domains of publication output, scientific collaboration and scientific impact. This chapter is thus directly concerned with the pathway from data quality to reliability of

bibliometric indicators.

Chapter 5, in a first part, continues the empirical study of bibliometric data quality by looking at WoS citation counts. The citation counts calculated by WoS are compared to citation counts derived from semi-automatic cited reference searches, that is, manually collected gold standard data. This chapter is therefore also concerned with the influence of data quality on the reliability of bibliometric indicators. In the second part, a new approach to the propagation of uncertainty from data quality to indicator values is introduced. This approach builds on Bayesian modelling of error distributions and posterior prediction of simulated error-free values. I show how this approach can use knowledge on statistics of inaccuracies from empirically collected samples to generate error margins for the values of bibliometric indicators of interest. The method is applied using both the citation count error data from the first part of the chapter and the data on document type misclassification from Chapter 2. It is noteworthy that this method can in principle be used for propagation of errors from any bibliometrically relevant data. The usefulness of this method in real bibliometric evaluations is shown for data from an evaluation of the research output of chemistry research groups. In this case study, I also use the month-precision citation windows from Chapter 3. This chapter builds on the results of the earlier ones but goes beyond them in sketching a method which allows to use the many empirical results on data inaccuracies to construct errors margins for bibliometric indicator values, an important demand of the Leiden Manifesto (principle 8).

In these studies, a number of different issues related to reliability and validity of bibliometrics are studied which are primarily related to data quality. In some cases it can be shown that the processing of the original data for the bibliometric databases is the location where inaccuracies are introduced (document type classification), in some cases data is missing that is in principle available (publication month data), in other cases the original data are heterogeneous, or unharmonized, and the processing by the bibliometric information systems only manages to resolve some, but not all, of this heterogeneity (cited reference strings, address affiliation strings). The common approach to quantifying the data quality in the bibliometric information systems is the direct empirical comparison with outside reference records. Besides identifying, quantifying, and in some instances demonstrating methods of remediating such errors and inaccuracies, one central concern of these studies is to investigate the effects of these

data quality issues on bibliometric practice in basic and applied research. The intention is to show that these are not mere technicalities or errors so small that they vanish in large scale data analysis. I am confident that after having read the next few chapters, the open-minded reader will agree that these problems do actually matter a great deal, that they deserve to be taken seriously and that the proposed solutions can contribute to improving the professional practice of science studies.

2 Document type assignment accuracy in the journal citation index data of Web of Science

This chapter was first published under the same title in *Scientometrics*, vol. 113, pp. 219–236, at <https://doi.org/10.1007/s11192-017-2483-y>, on 04 August 2017.

Abstract

This article reports the results of a study of the correctness of document type assignments in the commercial citation index database Web of Science (SCIE, SSCI, AHCI collections). The document type assignments for publication records are compared to those given on the official journal websites or in the publication full-texts for a random sample of 791 Web of Science records across the four document type categories articles, letters, reviews and others, according to the definitions of WoS. The proportion of incorrect assignments across document types and its influence on document specific normalized citations scores are analysed. It is found that document type data is correct in 94 % of records. Further analyses show that within records of one document type as assigned in the data source, the records assigned to the type correctly and incorrectly have different average page counts and reference counts.

Keywords

citation normalization; document type; data accuracy; bibliometric data; citation impact; Web of Science; Scopus; data quality

2.1 Introduction

Scientific journal publications can be classified by their intended information transfer function into distinct document types (also referred to as DTs in the following) such as research articles, reviews or letters. The DT data in citation index databases contain useful information about scientific publications that is routinely used in bibliometric studies.

The specific importance of DT in bibliometrics lies in their usage for 1) defining included and excluded publications in analyses and 2) in creating appropriate reference sets for calculating expected values for normalized citation scores and, less commonly, 3) in weighting publications according to their DT.

Any studies making use of the document type data have so far implicitly assumed that the document type data of commercial citation index databases is sufficiently accurate for these purposes. However, to the knowledge of the author, there is no published study in which this assumption was tested as of yet.

The reason for making distinctions across publications according to their DTs in scientometrics and research assessments is that because of their specific purposes and contents, they are utilized, that is, read and cited, differently, which leads to different citation distributions (e.g. Braun, Glänzel, & Schubert, 1989b, p. 392, Table 6; Lundberg, 2007, p. 156, Fig. 2; Vinkler, 2010, p. 176 ff.). In a thematically homogeneous set of publications, reviews are cited on average more often than articles, which are cited more often than letters (Glänzel, 2008, p. 14). Publications of different document types also exhibit different onset and decay in the distributions of citations over time, as well as in average “citation speed” (J. Wang, 2013). This leads to very different results in calculations of citation-based indicators when constructing reference sets by DT and when disregarding this kind of normalization (Moed & van Leeuwen, 1995; Sirtes, 2012). Therefore, when computing relative indicators of citation impact, the correctness of the assignment of document types to publications is crucial for fair comparisons and valid results. Furthermore, as DT is also commonly used to restrict publication sets from bibliometric datasets to those containing primary research results, DT data quality is highly important for obtaining correct publication counts for research units. As an example, in a comparison of pharmacology and pharmacy journals indexed in Web of Science and Scopus, Gorraiz & Schloegl (2008) found that in one fifth of the titles, the sum of articles and reviews differed in the two databases by more than 10 %.

There is no common standard of DTs in scientific publishing. Publishing houses, journals, and databases all use their own nomenclatures and definitions. The producers of Web of Science and Scopus do not directly copy the DT information of journals. Both producers maintain fixed systems of DTs for use in their products which can be accessed in the respective database documentation¹². How the internal DTs are exactly assigned based on these systems is not appar-

¹²For Web of Science: Web of Science® Help. Searching the Document Type Field. [accessed 2016/10/07] http://images.webofknowledge.com/WOKRS59B4/help/WOS/hs_document_type.html. For Scopus: Scopus Content Coverage Guide Jan. 2016; page 10. [accessed 2016/10/07] https://www.elsevier.com/___data/assets/pdf_file/0007/69451/scopus_content_coverage_guide.pdf

ent from the documentations. The correctness of DT assignment in commercial citation indexes has been called into question. Van Leeuwen et al. (2007) draw attention to the treatment of letters and ‘research letters’ from medical journals that were considered to be the same type in Web of Science (WoS). Spodick & Goldberg (1983) categorized samples of letters to the editor in four general and four specialist medical journals, published in 1979, according to their function. In the specialist journals these were comprised mainly of letters concerning articles published earlier (79 % of letter items). Very few items were replies to such letters. Letters in response to editorials and “letters presenting initiatives”, which often reported primary research in brief form, were also found. In the general journals as much as one quarter of letters was of the type “presenting initiatives”. These results indicate that a sizable proportion of items in letter sections do not address previously published research or continue such discussion through author replies. Across the eight journals it was found that within letters concerning published articles, 11-45 % of the items were criticism, 4-45 % were replies to such criticism and 5-22 % were supportive or supplemented the findings.

A similar study of Tierney et al. (2015) scrutinized the roles of letters in four otorhinolaryngology specialty journals published in 2012 (92 letter items). Responses to original articles accounted for 33 % of letters, author replies to them accounted for 20 %. Responses to letters made up 9 % and letters not related to original journal material 38 % of the total. For example they find:

Clinical Otolaryngology [a journal] was the most varied when it came to letter theme with 55% being unrelated to original journal material, of which technical innovations were to the fore, representing 13.2%. 9.4% (n = 6) were pilot studies, with literature reviews representing 7.5% (n = 4). Other categories’ letters fell into were spontaneous discussion/educational topics, new tools and case reports. (Tierney et al., 2015)

These two studies illustrate that the letters sections of medical journals contain a variety of material, a part of which is discussion of articles, a significant part is wholly different in content.

Another point of contention is the correction of assignment of review articles. Sigogneau (2000) and Harzing (2013) illustrate how WoS was using some highly

questionable assignment criteria for classifying publications as reviews in the past. Another case in point are the discrepancies in the number of items of DT article and review found by Gorraiz & Schloegl (2008) for the same journals across the two investigated data sources. As DT data is important in practice and its accuracy has been questioned in the literature, there is a need for an assessment of the correctness of DT data in citation index databases.

Besides document type errors, there are numerous other classes of error in bibliometric databases which threaten the validity of results obtained from these databases. To name but a few striking observations discussed recently in the literature, Valderrama-Zurián et al. (2015) found a surprisingly high incidence of duplicate publication entries in Scopus. Extensive studies on omitted citations have been performed by Franceschini and colleagues. Using a method that uses WoS and Scopus data in parallel (Franceschini et al., 2013) they obtain rates of missing citations on the order of approximately 5 % in these sources. This allows them among other things to indicate the accuracy of journal impact indicators (Franceschini et al., 2015b). In the same paper it is shown that journal omitted citation rates not infrequently reach figures above 10 %, in the extreme even up to 40 %, in the studied field of manufacturing. They furthermore report large numbers of duplicate DOIs, which are commonly used to match external records to bibliometric database items (Franceschini et al., 2015a). Further evidence is provided by Gorraiz & Schloegl (2008), who noted large discrepancies of the number of articles and reviews of the same journals indexed in both WoS and Scopus.

The overall impression of considerable negligence of data quality increases concerns about the basic validity of any bibliometric study using these data sources, independent of disputes regarding conceptual foundations.

In this article, DT assignments for a sample of WoS records are compared to those of the publishers on official journal homepages and in the original publications' full-texts, using WoS' own definitions of DTs. We also compare the DT assigned in WoS to the DT assigned in Scopus to the same publications by linking the two datasets for the WoS-sampled records. As we found a considerable proportion of disagreements between our independently coded DT and those of either data source, we also investigate the effect that this data inaccuracy will have on typical bibliometric studies. To this end we calculate two field normalized citation scores for each publication with an incorrect DT, one based on the DT as available in the data source and one with the corrected DT. We use the

distribution of score differences to shed light on this possible distorting source of error. By reporting on the error distribution across the four DT classes, we hope to enable researchers and practitioners in bibliometrics to better judge the influence of data inaccuracy on their results.

2.2 Methods

2.2.1 Data and Sampling

For this study, data licenced from Thomson Reuters¹³ Web of Science (WoS) and Elsevier's Scopus and loaded into custom SQL databases was used. The WoS data is comprised of the collections Science Citation Index Expanded, Social Sciences Citation Index and Arts and Humanities Citation Index and is the primary object of study and the data source actually sampled. The Scopus data linked to the sampled records is used only in a supportive role for some comparisons but is not sampled and hence not studied in depth. The WoS sample was restricted to publications published in journals (no proceedings papers, books or book chapters were considered), and restricted to publication dates between 2002 and 2014. The distribution of WoS-assigned DT in journals for that period is: articles: 67.6 %, letters: 2.6 % reviews: 3.5 %, others: 26.3 %. The third restriction is that only those records that had an assigned DOI in WoS were retained in the sample. The reason for this is that the DOI is used as one of several data fields for linking WoS records to Scopus records to make comparisons across data sources feasible. As a consequence, only relatively recent publications are sampled, see Appendix B for the distribution of publication years of the sampled records. There is one complication to this. It was reported that DOI in citation indexes are not as unique as they should be by definition (Franceschini et al., 2015a). For this reason, before the sampling was carried out, both databases were checked for DOIs that occur more than once and the combined list of those DOI was used as a further exclusion criterion for the sample so that no duplicate DOI would lead to ambiguities in the matching step¹⁴. The population size resulting from these restrictions is approximately 11.07 million records.

¹³The department of Thomson Reuters producing the citation indexes recently changed ownership and now runs under the company name Clarivate Analytics.

¹⁴For our WoS data (publications from 1980 to 2014) this list of multiply assigned DOI contains 19,096 entries and the list for Scopus data contains 81,715 entries (publications from 1996 to 2014).

Random samples of document identifiers were drawn from the WoS database, stratified by document type as assigned in WoS after four document type groups were formed. The groups are the types *article*, *review* and *letter*, which were used directly as available in the data, as well as records not assigned to one of those three types, in this study referred to as *other*. Subsamples of the document types were drawn in approximately equal numbers, using the WoS-assigned DT. It was attempted to evaluate approximately 200 publications of each of those four categories to get a reasonably sized sample with subsets of similar size. This allows for an overall estimate of accuracy with a 5 % margin of error. It is necessary to oversample the more rare DTs to be able to make accurate estimates of the error rate within these subsamples. Stratification by WoS DT was used because articles account for the vast majority of all publications and we were interested in estimates of the correctness of DT assignments of the population and the four relevant DT class subsets. The grouping into articles, letters, reviews and others follows the convention of distinguishing between primary research results (articles), secondary, synthesizing overviews (reviews), discussion about published results (letters) and others, such as editorial content or book reviews. The latter are usually deliberately excluded from bibliometric analyses, because they do not relate closely to research. This is necessarily a simplification. Editorials and book reviews can in some contexts be considered items relevant for bibliometric studies (Van Leeuwen et al., 2013; Zuccala & van Leeuwen, 2011).

The sampled records were linked to the Scopus records detailing the same documents using a matching table constructed for this purpose. The matching procedure is described in Appendix A.

To summarize the sampling process:

1. restrict WoS data to journal publication records
2. restrict to records from 2002 onwards
3. restrict to records with unique DOI
4. independently sample the four WoS/synthetic DT categories article, review, letter and “others” (not of the former three types) on the order of 200 records each.

2.2.2 Data collection

Bibliographic description data of each sample record comprised of article title, first author family name and initials, publication year, journal name, volume, issue and first page were queried from the WoS data and saved along with internal record IDs into a separate table. The table rows were randomized and the data exported as a spreadsheet file. The author and one student assistant independently searched for the publications' abstract web pages and full-texts online, using the bibliographic information to query academic and web search engines, blind to the DT in the original data. On the article web page of the journal or in the full-text, we attempted to find the officially assigned document type, if specified, and initially coded it as *article*, *letter*, *review*, *other*, *document not found*, or *ambiguous*. The individual DTs collected into the synthetic "other" category in this study are not distinguished as they are of only minor interest to bibliometrics, which is primarily concerned with publications providing scientific results. If no document type was stated, all useful information such as title, abstract and full-text was taken into account and used to decide on the DT to be assigned. We did not rely on journal section headings exclusively, only in combination with other evidence. The source URL and the term used to describe the DT in the source were recorded. When we found an article from the field of medicine, we also recorded the fact whether the publication is a case report or not.

The definition of document types used for the independent assignment exactly followed those of the two data source providers, which are sufficiently similar on the level of the four classes used. To summarize, an *article* is a report on original research of any length, including meta-analyses. Proceedings papers published in journals are also included in this class. A *review* condenses many previously published original research findings on a specific topic. A *letter* is a correspondence with the editors and readers of a journal about an item previously published in the same journal. The special rest category *other* comprises everything else, for example editorials, book reviews, meetings abstracts and corrections. This includes comments on papers published in different journals and invited discussion contributions that follow focus articles.

After the results were obtained the assignments of the two raters were compared side by side and any disagreements and remaining uncertainties were resolved by reviewing any available information and further searching for clarifying in-

Table 2.1: Sample description - DT frequencies

DT	independent assignment	WoS	Scopus
Article	294	204	279
Letter	111	193	177
Review	165	185	166
Other	208	209	100
Hybrid	13		
Total	791	791	722

Note: document type category “Hybrid” only applies to the results obtained by independent assignment

formation when necessary, including all cases when at least one rater indicated a record as *not found* or *ambiguous*.

While initially a total of 793 publication records were assessed, two records had to be excluded from the analysis because no certain DT assignment was possible as no reliable information for the publications could be found. This resulted in a final sample of 791 cases. The basic properties of the sample are described in Table 2.1 and Table 2.10 (Appendix B). In 13 cases we found it not to possible to assign the publication to strictly one of the DT categories exclusively. Instead, they appeared to be hybrids between two types. Consequently, it was decided that they will be recoded to have two permissible DTs, such that they are counted as correct in the comparison if the DT in the citation indexes is one of these two DTs. These publications comprise one “article/letter”, 4 “article/reviews” and 8 “article/others”. Of the final sample, 722 records (91 %) could be matched successfully with Scopus records.

2.2.3 Data analysis

The collected data was compared with the DT assigned by WoS and Scopus, respectively, and agreement was coded as 1, disagreement as 0. The transformation of WoS DT classes into our four relevant DT classes was straightforward. In the case of Scopus data, records coded as “short survey” were recoded as *reviews* (15 cases), as their description in the Content Coverage Guide explains that they are a shorter type of review. Records coded as “proceedings paper” that were published in regular journals are recoded as *articles* (19 cases) and

one record coded as “in press” was recoded as an *article*.

The data quality of the DT field was assessed by the share of correct DT assignments in the databases. Weighting by population proportions on the DT strata (as coded in WoS) was used for statistical inference from the sample to the population which is possible only for WoS, being the source of the sample. The population frequencies for the values of the variable are known from the original data. Overall and DT category inferential statistics were calculated with the *survey* package for the R statistical programming environment, version 3.3 (Lumley, 2004).

The measure of Precision, as used in information retrieval evaluation, serves as the main indicator of DT assignment correctness in this study. In information retrieval, Precision is defined as the proportion of retrieved documents that are relevant to a query among all returned documents. Analogously, Precision is understood in this study as the proportion of correctly assigned records among all records assigned a specific DT and can be calculated by the following general formula (Baeza-Yates & Ribeiro-Neto, 1999, p. 75)

$$P = \frac{\text{number of relevant records retrieved}}{\text{total number of retrieved records}},$$

or in the terms of this study

$$P = \frac{\text{number of correctly assigned records of a specific DT}}{\text{total number of records assigned to a specific DT}}.$$

In the case of Precision calculation only, hybrids are also counted as correct. The applied method of calculation of the strata weighted confidence intervals follows Korn & Graubard (1998). To estimate the proportion of correctly coded DTs in the population across all DT categories it is necessary to take into account the stratified sampling design. This is accomplished by defining the sampling design using the four strata groups and their population proportions in the *survey* R package and then computing the overall mean of agreement between database-assigned and independently assigned DT.

It is similarly possible to estimate population values for the other main quality criterion from information retrieval, Recall. In information retrieval, Recall is the proportion of retrieved documents that are relevant. In this study this is equivalent to the proportion of correctly assigned records among all records that

actually have that DT. If there are assignment errors, those records cannot be ‘retrieved’ under their correct DT. The basic formula for Recall is

$$R = \frac{\text{number of relevant records retrieved}}{\text{total number of relevant records}},$$

which in terms of this study becomes

$$R = \frac{\text{number of correctly assigned records of a specific DT}}{\text{total number of records of a specific DT}}.$$

For Recall calculation, 14 records assigned as hybrid DT were excluded, as there is no sound way to use them. It is not appropriate to directly calculate *sample* Recall values in this study, because they have little relevance in light of the uneven population proportions of DT categories as opposed to the approximately equal strata sample sizes based on WoS DTs. Since records were sampled with probabilities different to their population frequencies, the calculation of the population Recall estimate needs to be weighted accordingly. The modified formula, following the method for estimation of population proportions from stratified sampled data (Lohr, 2010, p. 80-81), is

$$R'_d = \frac{\pi_{d,d} \times w_d}{\sum_{i \in D} \pi_{d,i} \times w_i},$$

where subscript d denotes a specific document type from the set $D = \{A, L, R, O\}$ of all document types (strata), $\pi_{d1,d2}$ refers to the proportion of records of DT $d1$ in the stratum of DT $d2$ and w_d is a weight derived from the population proportion of a DT d . For instance, the Recall for article DT records in WoS is calculated this way:

$$\begin{aligned} R'_A &= \frac{\pi_{A,A} \times w_A}{(\pi_{A,A} \times w_A) + (\pi_{A,L} \times w_L) + (\pi_{A,O} \times w_O) + (\pi_{A,R} \times w_R)} \\ &= \frac{\frac{197}{204} \times 0.675}{(\frac{197}{204} \times 0.675) + (\frac{69}{193} \times 0.026) + (\frac{10}{203} \times 0.263) + (\frac{19}{181} \times 0.035)} = 0.96 \end{aligned}$$

For a simplified illustration of the calculation please refer to Appendix C.

2.3 Results and discussion

Qualitative findings

There is no standard for the designation of document types in scientific publications. Many different designations for the same DT were observed on different journal web pages and publication full-texts in collecting the data. This creates ambiguity, as the following examples of alternative designations which were found for articles, reviews and letters illustrate.

- Article: original research article, short report, short communication, full paper, original investigation, rapid communication, communication, progress report, case study, brief report, research letter, method article, clinical report, letter (“reports an important novel research result, but is less substantial than an Article” in the journal *Nature Biotechnology*)
- Review: mini review, critical review, review article, survey
- Letter: comment, correspondence, letter in reply, communication to the editor, response, discussion paper (“critical comments on papers already published in the Journal” in the journal *Electrochimica Acta*)

This list is not exhaustive and many variations of the terms occurred. The examples also show that merely using the term as given by the journal is not sufficient as these can be ambiguous and are not standardized, which is in line with the conclusions of Montesi & Mackenzie Owen (2008).

The publication of short original research findings and case reports as “research letters” addressed to the journal editors is common in medicine and closely related disciplines, confirming the findings of Van Leeuwen et al. (2007). Because they report original research, they are considered *articles* in this study. This follows directly from the official definitions of the two data sources in their documentations, which state that case reports are considered as articles in terms of their DT. Furthermore, many journals make distinctions between articles based on publications of completed research and manuscripts presenting preliminary results or specifying different DTs simply based on the length of the paper (“full length paper” – “brief communication”). A similar distinction was also observed for reviews in some cases (“full review” – “mini review”). Special types of articles are often treated and flagged separately, such as case studies, meta-analyses, methodology articles and tutorial papers. This information is useful for bibliometrics (Barrios et al., 2013; Patsopoulos et al., 2005; Romero et al., 2009). For example, case reports make up a significant fraction of what are presently coded as *articles* in medical journals. If they were identified distinctly it would

Table 2.2: Independent DT assignment vs. that of WoS

independent assignment	WoS				Total
	Article	Letter	Other	Review	
Article	197	69	10	19	295
Letter	1	105	5	0	111
Other	1	17	185	5	208
Review	5	0	3	157	165
Article/Letter Hybrid	0	1	0	0	1
Article/Other Hybrid	0	2	6	0	8
Article/Review Hybrid	0	0	0	4	4
Total	204	194	209	185	791

Note: Agreement between independently assigned document type and database assigned document type is indicated in bold

be possible to attempt to reduce noise in indicators by excluding them from analyses or treating them separately, as they are rarely cited (Patsopoulos et al., 2005). In fact, treating them separately might be appropriate as there are a number of journals which only or mostly publish case reports. At present, this is possible by using PubMed data, but not by using data from WoS or Scopus only.

Descriptive results

The complete result for the independent assignment compared to those of WoS for the same publications is given in Table 2.2 and the result for Scopus in Table 2.3. In the case of the WoS data, 655 of 791 cases were assessed as correct (83 %). For Scopus the figures are 549 correct cases out of 722 (76 %). It is important to note that no estimate of the Scopus population values can be derived from this because only WoS was sampled and the two data sources differ in the inclusion of content types. For the WoS sample we note that according to the data source's DT assignments, there were 194 letters, while we found only 111 plus one article/letter hybrid. WoS data indicates 204 articles in the sample while we found 295 plus 13 hybrids.

Case reports

As noted before, many medical case reports appeared in the letters section of issues or were addressed directly to the editors. This might lead to them being incorrectly assigned the DT letter. In the documentations of both Web

Table 2.3: Independent DT assignment vs. that of Scopus

independent assignment	Scopus				Total
	Article	Letter	Other	Review	
Article	218	59	3	10	290
Letter	4	99	7	0	110
Other	27	16	88	18	149
Review	23	0	2	135	160
Article/Letter Hybrid	0	1	0	0	1
Article/Other Hybrid	4	2	0	2	8
Article/Review Hybrid	3	0	0	1	4
Total	279	177	100	166	722

Note: Agreement between independently assigned document type and database assigned document type is indicated in bold

Table 2.4: DT assigned to case reports in WoS and Scopus

independently assigned DT	DT in data source	frequency WoS	frequency Scopus
Article	Article	3	11
Article	Letter	33	33
Article	Other	6	1
Article	Review	4	1
Article/Other Hybrid	Article	0	1
Article/Other Hybrid	Other	1	0

Note: Agreement between independently assigned document type and database assigned document type is indicated in bold

of Science and Scopus case reports are explicitly stated to be assigned the DT article. In the WoS sample, there were 47 case reports. The majority of them were assigned incorrectly: 33 letters, 4 reviews, 7 other; 3 were correctly coded as articles. Hence this is a major source of error but does not account for all the inaccuracies in records coded articles by WoS. All 47 case reports are also contained in Scopus, where the error is not as large: 12 records are correctly identified as articles while 33 were coded as letters. One case report fell under the DT review and one under the DT other. The findings are summarized in Table 2.4.

Comparing WoS and Scopus results directly

A cross-tabulation of the DT assignments of WoS and Scopus for the same pub-

Table 2.5: Comparison of DT assigned by WoS and Scopus to the same articles (N=722)

according to WoS	according to Scopus			
	Article	Letter	Other	Review
Article	191	1	1	9
Letter	11	174	5	0
Other	38	2	93	19
Review	39	0	1	138

Table 2.6: Comparison of DT assigned independently and by WoS for publications not included in Scopus

independent assignment	according to WoS			
	Article	Letter	Other	Review
Article	2	1	1	0
Letter	0	1	0	0
Other	0	1	56	2
Review	0	0	0	5

lications is given in Table 2.5. The two databases give the same DT, according to our four categories, in 83 % of the cases in this sample (N=722). Notable is the discrepancy in what Scopus considers *articles*.

It could be possible that the differences in coverage of the data sources influence the observed disagreements. To show that this is unlikely, Table 2.6 shows the independently assigned DT and the DT in WoS for publications found only in WoS but not in Scopus. No systematic pattern is apparent. The high number of records of type *other* not found in Scopus is consistent with Scopus’ policy of not including meeting abstracts, which WoS does cover.

Statistical analysis

The statistical inference is restricted to WoS as a data source, as only WoS was sampled, while Scopus data was subsequently linked to these sampled records. The coverage of WoS and Scopus differ, as does their DT inclusion policy. Hence, no reasonable statements of DT assignment accuracy for Scopus can be made from this data.

Table 2.7 contains the results for the estimation of Precision across DT classes

Table 2.7: Population estimates for Precision in DT assignment in WoS (n=791)

DT category	Web of Science
	Precision (95 % confidence interval)
Article	0.97 (0.93, 0.98)
Letter	0.55 (0.48, 0.62)
Other	0.91 (0.87, 0.95)
Review	0.87 (0.81, 0.91)
total	0.94 (0.92, 0.96)

Table 2.8: Population estimates of Recall in DT assignment in WoS (n=778)

DT category	Web of Science
	Recall (95 % confidence interval)
Article	0.96 (0.95, 0.98)
Letter	0.58 (0.36, 0.79)
Other	0.97 (0.95, 1.00)
Review	0.57 (0.39, 0.75)
total	0.94 (0.92, 0.96)

for WoS. The population estimate of the WoS overall proportion of correctly assigned DT is 0.94; 95 % confidence interval: (0.92, 0.96). The results for the estimation of DT class and overall Recall values are given in Table 2.8.

To summarize the results, Precision is particularly low in the DT *letter*. In terms of Recall, the assignment quality for the DT *article* and the synthetic DT of *other* are high and for *letter* and *review* the values are low. A considerable part of the inaccuracies in the DT *letter* results from incorrectly assigned case reports. Even with manual coding effort there remains a small set of publications which cannot easily be assigned to exactly one category.

Differences in quantitative publication characteristics between correctly and incorrectly DT coded publications

Because of their different communicative purpose and requirements for authoring, publications of different document types have specific distributions of various characteristics. For example, articles are usually longer than letters and reviews usually have more references than articles. For the WoS data collected in this study, the mean of the distributions of the number of pages, number

Table 2.9: Quantitative characteristics of different DT correctly and incorrectly assigned (WoS data)

DT in WoS	correctly assigned	N	mean pages	mean authors	mean references
Article	no	7	10.9	2.4	49.7
	yes	197	8.9	4.5	31.9
Letter	no	87	2.7	4.3	7.6
	yes	106	1.7	2.2	5.8
Other	no	18	3.8	2.9	15.7
	yes	191	2.2	5.9	5.1
Review	no	24	8.1	3.6	47.1
	yes	161	13.1	3.1	99.7

of authors and number of references are given in Table 2.9 for correctly and incorrectly coded DT assignments, respectively. The data confirm the above assumptions: the averages differ notable for correctly and incorrectly assigned records. This suggests the possibility of using these and possibly additional data to find publications which are likely coded incorrectly, namely those that deviate in such characteristics from the expectations of the typical publications of that DT. While we did not attempt to exploit this possibility in this study, this observation suggests the usefulness of a semi-automatic method to reduce the needed manual checking in situations where the DT data correctness is particularly important.

Influence on citation indicators

It was found that a certain share of DT assignments is inaccurate. Insofar as the uncorrected DT data are used to define reference sets to compute normalized citations scores (NCS), the data inaccuracies will influence citation impact studies. Impact distortions occur because the observed citation count is divided by the wrong DT-based reference group’s average citation score (expected value). Large-scale assessments that use NCS with normalization by DT include the Times Higher Education World University Ranking¹⁵ and the National Science Foundation’s Science and Engineering Indicators 2016¹⁶.

¹⁵Documented for the 2015/2016 edition <https://www.timeshighereducation.com/news/ranking-methodology-2015> but not the 2016/2017 edition <<https://www.timeshighereducation.com/world-university-rankings/methodology-worlduniversity-rankings-2016-2017>>.

¹⁶<https://www.nsf.gov/statistics/2016/nsb20161/#/>

To study the extent of this effect, the normalized article citation scores for 3-year citation windows for publications that were assigned an incorrect DT in WoS were computed. We only included records with both a corrected DT and a WoS DT of article, review or letter that were published prior to 2012 (73 cases). The calculation follows the method described in Waltman et al. (2011). WoS Subject Categories were used for disciplinary aggregation of publications necessary for defining the reference set. The incorrect values are the NCS computed with the reference set citation rate averages¹⁷ for the WoS-assigned DT while the corrected NCS are computed using the reference set average citation rate of the actual DT of the publication, as determined by this study, see Fig. 2.1.

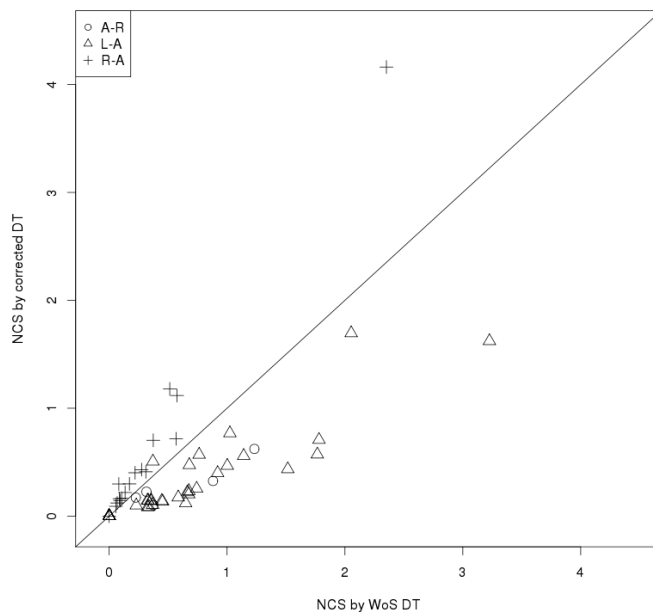


Figure 2.1: Scatterplot of WoS and corrected normalized citation scores for incorrectly coded publications ($N=73$). *Circle*: corrected from article to review (5 obs.); *triangle*: corrected from letter to article (51 obs.); *cross*: corrected from review to article (17 obs.)

The figure shows that these publications' NCS are over- or underestimated, de-

¹⁷It should be kept in mind that all the expected values are themselves affected by DT inaccuracies. No corrected values for these reference scores are available, so this issue must be put aside for this study although it is relevant in general.

pending on the specific pair of incorrect and correct DT. The diagonal indicates the line of no error. Publications with correctly computed values would lie exactly on the diagonal. Consider for example the publications whose DT was corrected from letter to article (triangle symbol). Most of them fall below the diagonal, indicating that their citation impact was overestimated because they were divided by expected values of letter publications even though they should have been divided by the expected value of articles. One relatively isolated data point is the cross symbol for an article mistakenly assigned the DT of review in WoS at the coordinates (2.4, 4.2). This publication, the paper Chaiworapongsa et al. (2008) is clearly an article, as the full text reports a full medical study and is labeled on the abstract page as “Research article”¹⁸ and on the issue contents web page it is listed in the section “Original Article”. It was cited 29 times in three years, while articles in its subject class were cited 7.0 times and reviews 12.3 times on average.

The error distribution of the NCS in terms of corrected scores minus WoS scores is shown in Fig. 2.2. 24 of the 73 publications have zero citations and are consequently not affected by distortion. It should be kept in mind that this histogram is not of the NCS error distribution of the entire sample but just the subset of incorrectly assigned DT.

These results confirm that within this specially selected subset of inaccurate DT data normalized citation scores of individual publications are strongly affected by these inaccuracies. Unless the citation count is zero the change in NCS is usually substantial and increases as the absolute citation count gets higher. The systematic influence of the kind of incorrectness, that is, which specific DT assignment error occurred, is also evident.

Limitations

This study has some methodological limitations that need to be considered when interpreting its results. The sample was restricted to publications that had a (unique) DOI assigned. In the earlier years of the sampling frame, this is not the greater part of publications. If the accuracy of DT assignment in citation indexes differs for publications without DOI, the sample would be biased. The reference DT was assigned clerically by two raters. Such ratings are subjective to some degree. However, in the case of ambiguous DTs in the original source,

¹⁸<http://www.tandfonline.com/doi/abs/10.1080/14767050701832833> accessed Sept. 21, 2016.

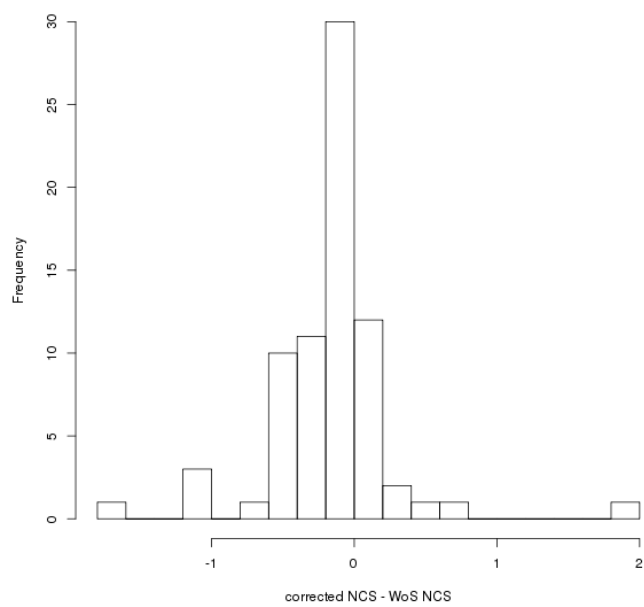


Figure 2.2: Error distribution NCS

some leniency was allowed for in the comparison. Because the sample is drawn from WoS data and afterwards matched with Scopus data where possible, the sample is not representative for Scopus and no estimates for the DT assignment accuracy of Scopus are possible.

2.4 Conclusion

In the present study, a stratified random sample of Web of Science-indexed publications was checked for accuracy of the document type as recorded in the data source in comparison to independently assigned document type data based on the data source's DT definitions. Document type assignments were found to be correct in Web of Science in 94 % of cases, when using the four DT categories of this study (article, letter, review, other). Significant proportions of publications may be missed when making selections of records from these data sources constrained to particular document types. In particular, records assigned the types letter and review are affected by inaccuracies. For identical publications indexed in both data sources, the DT assigned by WoS and Scopus are often conflicting. Because of different citation distributions of publications of different document types, the normalized citations scores of publications with incorrect DTs become systematically distorted.

In analyses of modest numbers of publications, publications should not be excluded from citation analysis a priori because of their apparent document type. DT assignments should be verified if feasible, especially for indicator supported research evaluation. When this is not viable, the results of studies can be accompanied by sensitivity analyses that take the information about inaccuracies of DT into account, using the error rate found empirically in this paper or determined independently for the publication set at hand. In interpreting the results of bibliometric studies, it should be kept in mind that when publications of some document types are excluded, this might lead to the unintentional exclusion of publications of interest. When citation normalization is based on reference classes of specific document types this can lead to inaccurate normalized citation scores. This problem is present in general in any bibliometric study that is based on calculations taking document type data into account implicitly or explicitly.

Acknowledgements

This study was supported by the German Federal Ministry of Education and Research (BMBF) grant 01PQ13001, project “Kompetenzzentrum Bibliometrie”. I want to thank Anastasiia Tcypina for help with data collection and Nees Jan van Eck for discussion of the manuscript.

Appendix

A: Matching of WoS and Scopus records

An initial basic matching table was created by FIZ Karlsruhe for general use in projects by the database users based on combinations of exactly corresponding values in 10 metadata fields. The fields were assigned weights for their discriminative power. For example, the publication year field has a low discriminative power while that of article title and DOI is very high. Field values were normalized for differences in character sets, capitalization, special characters and some structural aspects (e.g. removing separating dashes in ISSN) to make them more uniform between the two data sources. All journal records are then mutually compared across all fields and a score is calculated from the number and weights of exactly matching fields. Only pairs with a predetermined threshold score are kept. Through this procedure, a pair of records may be included multiple times in the resulting table because all reasonable field combinations are used. Furthermore, a record from one data source can occur as a plausible match pair with multiple records from the other data source. The matching quality of this method was assessed before the study was conducted and found to be satisfactory. For this, a random sample of 2450 matching pairs was selected from the table and manually checked for equivalence using full bibliographical data from both data sources. In 9 cases the match was incorrect, in 15 cases the decision was unclear and in 2426 cases the match was found to be correct, the percentage of correct matches being greater than 99 %.

This basic data was slightly modified for matching the sampled WoS records to those in Scopus uniquely. For WoS records being assigned only one possible Scopus record, the entry rows were copied directly to the final matching table. For those WoS records with more than one plausible match, the single row with the highest matching score was copied. This concerns 2.6 % of the entries in the base table. Using this look-up table, of the 793 records that were initially assessed, it was possible to identify matches in Scopus in 711 cases. This was

extended by matching only on the DOI for remaining records and manually assessing the matches. Finally, for all unmatched records that were left, the title was searched in the Scopus online platform and the results checked for a correct match. These two steps produced another eleven verified matches. The remaining unmatched records are mostly meeting abstracts records, which are deliberately not included in Scopus. According to the DT assigned for this study, these unmatched records comprise, after exclusion of publications which could not be found, four articles, one letter, five reviews and 59 others. Within these, there were five misclassifications by WoS.

B: Distribution of publication years in the WoS sample

Table 2.10: Distribution of publication years in the WoS sample

publication year	cases
2002	18
2003	31
2004	35
2005	42
2006	55
2007	60
2008	63
2009	79
2010	99
2011	90
2012	108
2013	93
2014	18

C: Equality of population estimates of Precision and Recall

In this study, the overall population estimates of Precision and Recall, as opposed to the estimates for particular DT, for each data source are equal, unlike in typical information retrieval evaluation. This is so because each case is “relevant” for one of the four DT categories. To illustrate, a simplified example is worked through.

Consider a dataset with only two document types, A and B, for which we have the data source’s assignments and the true DTs, as assigned manually. There

is no sampling and stratification. Data are cross-tabulated data source and independently assigned DT counts like this:

Table 2.11:

independently assigned DT	data source DT	
	A	B
A	20	6
B	5	40

There are 26 true A records, they have a population proportion of 0.37 while the 45 B records have a proportion of 0.63. Calculating first the DT Recall (R) and Precision (P) values for A and B gives:

$$R_A = \frac{20}{20+6} = 0.77 \quad R_B = \frac{40}{40+5} = 0.89$$

$$P_A = \frac{20}{20+5} = 0.80 \quad P_B = \frac{40}{40+6} = 0.87$$

Then the overall Recall and Precision are:

$$R = (R_A \times w_A) + (R_B \times w_B) = (0.77 \times 0.37) + (0.89 \times 0.63) = 0.85$$

$$P = (P_A \times w_A) + (P_B \times w_B) = (0.80 \times 0.37) + (0.87 \times 0.63) = 0.85$$

3 Effect of publication month on citation impact

This chapter was first published under the same title in *Journal of Informetrics*, vol. 12, iss. 1, pp. 330-343, at <https://doi.org/10.1016/j.joi.2018.01.012>, in February 2018.

Abstract

A standard procedure in citation analysis is that all papers published in one year are assessed at the same later point in time, implicitly treating all publications as if they were published at the exact same date. This leads to systematic bias in favor of early-months publications and against late-months publications. This contribution analyses the size of this distortion on a large body of publications from all disciplines over citation windows of up to 15 years. It is found that early-month publications enjoy a substantial citation advantage, which arises from citations received in the first three years after publication. While the advantage is stronger for author self-citations as opposed to citations from others, it cannot be eliminated by excluding self-citations. The bias decreases only slowly over longer citation windows due to the continuing influence of the earlier years' citations. Because of the substantial extent and long persistence of the distortions, it would be useful to remove or control for this bias in research and evaluation studies which use citation data. It is demonstrated that this can be achieved by using the newly introduced concept of month-based citation windows.

Keywords citation impact; publication date; citation analysis; bias; bibliometric indicators; self-citations; control variable; systematic error; validity

3.1 Introduction

Citation impact normalization is a central concept for the construction of advanced bibliometric indicators which eliminate the effects of different scientific discipline, type of document and date of publication (Waltman, 2016). By delineating sets of publications that are similar to each other in content and formal characteristics and by using these sets to compute reference values and by computing relative impact indicators based on these reference values, the heterogeneity in citation counts due to these factors is removed. The intention is to make possible fair comparisons, to compare like with like (Schubert &

Braun, 1986). The basic formal characteristics, as opposed to the content (disciplinary area), that are taken into account are document type (such as research articles, review papers, letters, editorials etc.) and publication date. Further characteristics have also been demonstrated to co-vary with citation counts, for example differences between methodological, theoretical and empirical works (Peritz, 1983), clinical vs. basic research in medicine (Van Eck et al., 2013) or clinical study level (e.g. Bhandari et al., 2007).

One important component of normalization is controlling for publication date, as, *ceteris paribus*, the more time has passed since publication the more papers will be published whose authors had the opportunity to read and cite a given publication. The publication year is commonly used to operationalize publication date. This practice is based on the implicit assumption that, for the question of interest of a study, it makes no difference when exactly in a year a paper was published. The fact that, when citations are counted at some later date, documents published in January have eleven months more to be read and cited than works published in December of the same years raises the question if the above assumption is justified, and, if it is not, under which conditions and how a more precise publication date ought to be used in citation analysis.

The question of the influence of a more exact publication date is related to the problem of choosing adequate citation windows, the period in which citations to papers in a set of publications are counted. A citation window that is very short, say two years, would more obviously lead to bias against papers published towards the end of the investigation period compared to those published towards the beginning. Consider the following simple illustration. Citations are counted at the end of the year after publication (2-year citation window). Then papers from January had 24 months to be read and cited, assuming the case they were published on the first day of the month and citations counted after the last day of the citation window, while December papers had 11 months less, just 13 months, which is 54 % of the time period of the January papers. This relative disadvantage becomes smaller as the citation window length is increased. In a five year citation window, for example, the December papers had 82 % of the citation duration of the January papers. Citations do not accumulate uniformly over time and one is not only concerned with January and December papers, so this reckoning does not say much about the actual size of the distortion. But it might serve as a first order approximation model. Just how big this ‘head-start’ effect is in reality and at what point in time it vanishes is the topic investigated

in this paper.

The article is organized as follows. In the next section, previous work on the topic is briefly reviewed and some knowledge gaps are pointed out which this study addresses. Next, the data on which the study is based are presented. The major part of this contribution is comprised of the analysis of the results regarding the month effect from several points of view, including its size as reflected in basic citation scores and in regression analysis, also taking into consideration the online publication date, the change of the effect size over longer citation windows and its presence and patterns across disciplines. Furthermore, we introduce a method to eliminate the month bias and use the resulting corrected citation counts to demonstrate the bias on a simulated academic impact assessment of institutions under realistic conditions similar to currently employed research evaluation procedures. We finish with a discussion of the results and their implications for the field.

3.2 Related work

There have been a number of prior studies that have noted and investigated the month effect. They will be briefly reviewed in the following and their results used as a point of departure for this study.

Haslam et al. (2008) used publication month as a control variable throughout their regression analyses of influence factors of citation impact in a psychology sub-discipline. Their criterion was the natural logarithm of articles' citations counted after ten years. In their results, the standardized regression coefficients for month are small, not statistically significant and of the expected negative direction in 3 out of 4 domain regressions and also negative in the final overall regression equation. The small number of publications studied (308) and long citation window might account for the inconclusive findings with regard to publication month influence.

Levitt & Thelwall (2011) proposed a new impact indicator based on a paper's early citation count and the impact factor of its journal and investigated the compound indicator's predictive validity. One of their research questions relates to the correlation between the proposed indicator and publication month and the correlation between raw citation counts and publication month. Their sample is comprised of UK articles in the Social Sciences Citation Index category

‘Economics’, published between 2001 and 2004 and they counted citations up to 2008. As an aside, the authors note that “the notional publication date of an article does not necessarily reflect the true publication date; frequently a journal issue appears many months before or after its notional publication date” (Levitt & Thelwall, 2011, p. 303). This applies to the present study as well and should be kept in mind as one of its limitations. The authors confirm the existence of the month bias of citation counts for their sample, finding advantages for early-month publications for up to 3 years after publication. The authors conclude that “it seems important to take into account month of publication when analyzing short citation windows” (Levitt & Thelwall, 2011, p. 307).

In a study of citation window lengths, J. Wang (2013) considered publication month as one of several factors possibly influencing citation ageing. Due to the design of the study, papers uncited after 31 years are excluded from the analysis. Wang also introduces a method for the estimation of publication month when this data is missing. The present study adopts this method, as will be detailed further below. The J. Wang (2013) study finds no discernible difference in citation ageing across months based on the first and last year after publication in which half or more of the publications are cited. While there is no month difference in ageing according to this measure, there are differences between research fields, document types and papers of different final citation count tiers. Another measure, ‘Citation Speed’, is also applied, which is based on the cumulative citation counts throughout the observation years. ANOVA analysis of this measure’s results does indicate significant differences in citation ageing by publication month, though the effect appears to be small.

Gai et al. (2015) study the month citation bias as such in one field, one specialty and one journal, respectively, for citation windows from 1 to 9 years. They exclude a few very highly cited papers. They group publication months into six two-month periods. For papers in physical geography they find significant group differences for citation counts for up to five years after publication. The results are similar for papers on the topic of diabetes. For the Journal of Biological Chemistry, group differences are significant up to the seventh year after publication.

The present study is intended to improve upon the outlined results in a number of ways. As previous studies have been limited to specific disciplines we investigate if the effect is general to all disciplines (as far as covered in Web of Science). Neither uncited nor very highly cited publications are excluded from analysis. A

range of disciplines with very different citation dynamics are investigated to provide insight into field differences. Field differences in publication month effect on citation counts can be hypothesized to mirror fields' size in terms of number of papers, their citation density (average references per paper) and publication and citation speed. Details on the rate of decrease of the month effect in relation to the length of citation windows have not been studied yet, so this issue will be given consideration. The present study takes into account citation windows of up to 15 years. The method for estimating missing month data proposed in J. Wang (2013) is validated. Month effects are measured for the two citation impact indicators average citations per paper and share of highly cited papers of single-months groups. The role of author self-citations as a specific factor of early-month citation bias is considered because self-citations generally occur earlier after publication than citations by others. The online publication date is tested for influence on citations in addition to that of the issue publication date. Finally, it is demonstrated how to eliminate publication month bias by constructing citation windows of equal length for month cohorts of papers.

3.3 Data and methods

3.3.1 Data sets

Data set A consists of all journal publications of document type 'Article' from the year 2000, obtained from Clarivate Analytics' Web of Science¹⁹ ($n = 767,959$), for which publication month data was either available in the source data or could be estimated, as will be reported below. This year was chosen in order to be able to follow the development of the month bias over long citation windows. 6539 articles from 2000 had to be excluded from analysis because they had neither a publication date value nor enough contextual data to obtain estimates of publication month. The number of citations to the articles in each year were computed from citation link data constructed using the 'iFQ' procedure, as described in Olensky et al. (2016), using references data from all journal publications of the period from the year 2000 to 2014. After a preliminary investigation of the possible effect of author self-citations it was decided not to remove them, as their influence on the sought after effect is very minor. This analysis can be found in the Appendix in Supplementary

¹⁹The data infrastructure of the German Competence Centre for Bibliometrics was used for data set A, cf. <http://www.forschungsinform.de/Bibliometrie/en/index.php?id=home>.

material. Disambiguated affiliation data are used for all publications with authors working in Germany for data set A (Rimmert et al., 2017; Winterhager et al., 2014).

Data set B is comprised of articles contained in PubMed, based on 2016 base files. Publications of the print publication year 2009 which had a DOI and online publication date available were selected and matched to WoS records based on the DOI. The records were restricted to the WoS document types ‘Article’ and ‘Review’. This results in 253,292 publications. Citation counts were computed by CWTS by their method which was also described in Olensky et al. (2016). Note that the two different reference matching procedures which were applied in the two data sets, give nearly the same results, while both improve upon the citation matching links supplied by the vendor (Olensky et al., 2016). Self-citations are not removed. Data set B will be used to assess the combined influence of online and print publication months, for which more recent data containing both print and online publication dates is required.

3.3.2 Missing month data

Publication month data is not available for all items in data set A. A publication date containing a specific month was present in 82 % of the records. These were used directly. A publication date span of more than one month or a quarterly designation was present in 7 % of the articles. For the former cases the numerical values were averaged, which is also considered part of the estimation in the following. For a number of journals appearing quarterly the publication date in the source data is one of spring, summer, fall or winter. By inspecting the order of issue numbers of quarterly journals with such seasonal publication dates it was found that a large percentage of titles does not follow the parallel issue order of 1-4 and spring to winter. On the basis of this inconsistency it was decided not to apply a simple uniform transformation of the season designation to numerical month values. Thus, for these quarterly journals and the remaining 11 % of the records without publication date information and it was attempted to estimate their month of publication on the level of journal issues.

One way to estimate the publication month is by utilizing the data available on the number and sequence of issues of the journal. The necessary assumptions are that volumes and issues are labeled sequentially in a year and issues are published in regular, equal intervals. These are simplifications as issues may

appear simultaneously or in irregular intervals. The publication month estimation used for this study follows closely the method of Wang (2013, p. 854). The integers from 1 to 12 are assigned to the months. Issues with a given publication date spanning more than one month are assigned the average numerical month, rounded to the whole number. For example, an issue with the publication date ‘OCT–DEC 2000’ is assigned the month value 11.

Publication months for issues without any publication date information or seasonal publication dates were estimated as follows. The number of issues of the journal in the year is counted. Issues are ordered by their volume and issue labels sequentially and assigned the resulting rank number. This rank is divided by the number of issues, multiplied by 12, and rounded to the nearest integer.

Table 3.1: Error distribution of month estimation

error (actual month – estimated month)	-10	-7	-6	-4	-3	-1	0	1	2	10
frequency	1	2	2	2	1	65	112	11	1	3

The accuracy of this estimation method was validated by (1) looking up the publication month(s) of the issue on the official publisher journal homepage. A random sample of issues without publication date information specific to a single month was drawn and the official publication date was looked up and recorded and subsequently compared with the estimated date. This includes quarterly publications, issues designated as spanning multiple months and issues without any publication date value. And moreover (2) by also applying the estimation method to issues for which the publication month was already available in the source data and comparing the source data, presumably correct, with the estimated data.

3.3.3 Indicators and statistical methods

Differences in the values of month-cohorts of publications are investigated for two citation impact indicators: average citations per paper (CPP) and share of highly cited papers. In the case of the latter, the share of papers of each publication month among the 10 % most highly cited papers across all months is tested. We assess how the impact values for each month cohort are statistically associated with the order of issue publication months with the Pearson correlation coefficient. In the case of the association of the month variable with

individual papers' citation counts, the correlation is also computed for the log-transformed citation count in order to minimize the effect of the non-normality of the citation distribution. To assess the compound effects of issue and on-line publication month on individual publication citation counts, we perform linear regression with log-transformed citation counts as the dependent variable (Thelwall, 2016).

3.4 Results

3.4.1 Missing month data estimation and validation

The month estimation method outlined above was applied for data set A to all issue records without publication dates specific to a month, that is, those with no data, with a range of months and with a seasonal date (i.e. 'SPR', 'SUM', 'FAL', 'WIN'). In order to validate the results, 339 issues were looked up on journal websites. For 199 of these issues the publication month(s) could be found. Issues spanning multiple months according to the official dates were transformed as described above. In the resulting data the estimated month correlates with the actual month of publication with a Pearson coefficient of $r = 0.908$ (CI: [0.88; 0.93]). There is an average difference of -0.44 months between estimated and true publication months, which means that the estimation results in a small underestimation; see Table 2.1 for the empirical error distribution. No pattern in the true publications months of the issues underestimated by one month could be detected. 94 % of the estimates in the sample had at most one month estimation error. This is accurate enough as a first-order approximation for the purposes of this study but can probably be improved.

Having established the validity of the estimation method with independently collected data we can furthermore use the already available publication date information for validation. To do so, we applied the estimation method to those issue records in data set A for which a specific single month is available in the original source data and calculated the correlation between actual and estimated values. We find for the 44,758 issue records a correlation of $r = 0.928$ (CI: [0.926; 0.929]). An accuracy of ± 1 month was recorded for 83 % of the issues.

As a third method of validation we compare the distributions of publications over months between estimated dates and publication dates given in the original data. In both cases it can be assumed that there will be roughly equal numbers

of publications in each month. That means that the theoretically expected distribution is a discrete uniform distribution with the values from 1 to 12 as the 12 events. This distribution would exhibit a mean of 6.5 and standard deviation (SD) of 3.45. The empirical figures for the articles with a single given publication month are: mean = 6.55 (SD = 3.45) and for those with estimated publication month: mean = 6.33 (SD = 3.45). Since the distribution of estimated months is in agreement with both the originally available data's empirical distribution and the theoretically expected distribution, the validity of the estimation method is further supported. It is noted that these figures are computed over all publications, not over the sample of issues as above. We stress again that both the nominal given publication month and the algorithmically estimated one may not be close to actual point in time when the issue could be read in full due to idiosyncratic publishing practices (cf. Garfield, 1978).

With this data at hand, it is possible to use descriptive statistics to show the numbers of papers published each month and analyze the shares of citations and ratios of citations per paper from single years or ranges of years for each publication month.

In data set A, the average number of papers per month is approximately 65,000 (SD = 4600), with a minimum of around 56,000 in July, and a maximum of around 72,000 in February. The minimum figure is about 78 % of the maximum. The number of publications per month is uncorrelated with month: $r = 0.02$, (CI: [-0.56; 0.59]).

3.4.2 Effect of publication month on citations at the level of month-cohort groups of papers

To examine the effect of the publication date on citations to papers published within one year, publications published in each month are considered as groups. The average numbers of citations per paper of all month groups are calculated in order to characterize the groups' citation impact. For purposes of exposition we examine first the average CPP per month based on citations from single years for the first four years before moving on to citation windows of multiple years. The results of monthly CPP for single years are shown in Table 3.2 and Fig. 3.1. For the first three years (2000–2002) the earlier months' publications receive on average more citations. For example, in the second year alone (the year following the publication year), January papers receive 2.08 citations while December

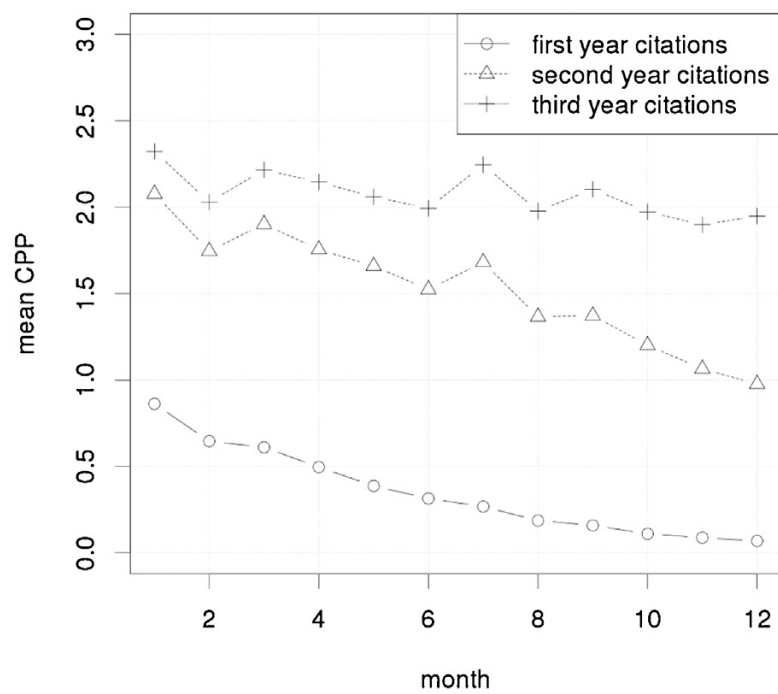


Figure 3.1: Average CPP per month for the citations of the first three years.

Table 3.2: Average CPP by month for individual years

month	citation window (year)			
	1	2	3	4
1	0.86	2.08	2.32	2.28
2	0.65	1.75	2.03	2.01
3	0.61	1.90	2.22	2.19
4	0.50	1.75	2.15	2.13
5	0.39	1.66	2.06	2.06
6	0.31	1.52	1.99	2.02
7	0.27	1.68	2.25	2.26
8	0.19	1.37	1.98	2.02
9	0.16	1.37	2.10	2.17
10	0.11	1.20	1.97	2.10
11	0.09	1.06	1.90	2.02
12	0.07	0.98	1.95	2.11
Pearson r	-0.97	-0.96	-0.69	-0.29
[confidence interval]	[-0.99; -0.88]	[-0.99; -0.86]	[-0.90; -0.19]	[-0.74; 0.34]

papers receive 0.98 on average, or about half as much. In the publication year and the second year the relationship between publication month and CPP is strongly linear negative (both $r < -0.9$). The association weakens in year three to $r = -0.69$ and for citations made in year 4 the effect is no longer present.

The strong negative correlations between publication month and average citations show that early-month publications enjoy a citation advantage for three years. Having established the effect for single citation years, we now move on to citation windows of *multiple years*. According to the above results, citation windows of two and three years after publication will necessarily exhibit the correlation. However, how long does it take for the effect of those early years to decrease so much to become negligible? In order to answer this question in the next step of analysis citations are counted cumulatively for subsequent years, which conforms to the most common practice in bibliometrics at present.

We consider selected citation windows of up to 15 years. The results presented in Table 3.3 indicate that the effect decreases rather slowly. For the five-year citation window, January papers outperform December papers by some 35 % in terms of average CPP. The correlation is as high as -0.75 after nine years, -0.68

after eleven years and at the end of the observation period, after 15 years, it is down to -0.57 .

Table 3.3: Average CPP by month for selected cumulative citation windows

month	citation window (year)						
	1-4	1-5	1-6	1-8	1-10	1-12	1-14
1	7.54	9.74	11.84	15.83	19.53	22.93	26.17
2	6.43	8.39	10.27	13.86	17.24	20.41	23.40
3	6.92	9.06	11.11	15.02	18.65	22.01	25.20
4	6.53	8.61	10.60	14.39	17.94	21.20	24.30
5	6.16	8.18	10.12	13.80	17.24	20.44	23.49
6	5.85	7.83	9.74	13.39	16.82	20.02	23.08
7	6.45	8.66	10.78	14.79	18.54	21.95	25.20
8	5.55	7.54	9.48	13.17	16.66	19.89	22.98
9	5.80	7.95	10.01	13.96	17.65	21.04	24.29
10	5.38	7.47	9.49	13.37	16.99	20.33	23.52
11	5.07	7.08	9.01	12.73	16.22	19.44	22.50
12	5.10	7.22	9.27	13.18	16.86	20.24	23.46
Pearson r	-0.92	-0.88	-0.83	-0.75	-0.67	-0.62	-0.57
[confidence interval]	[-0.98; -0.73]	[-0.97; -0.61]	[-0.95; -0.50]	[-0.92; -0.30]	[-0.90; -0.16]	[-0.88; -0.07]	[-0.86; -0.01]

It could be possible that the indicator CPP is affected more by publication month than other impact indicators because citation count distributions are very skewed and the mean of this distribution (CPP) is less robust than indicators based on other characteristics of the distribution. The averages of the citation counts of the month cohorts might be distorted in this way. To rule out such an effect, the publication month bias is also assessed with another frequently used impact indicator, the share of highly cited publications. We consider the share of each month’s publications among the approximately 10 % most highly cited publications.

In a publication set not influenced by publication month bias in citations counts, approximately equal shares of highly cited publications are expected to be present in each month group. The issue of ties, publications having a citation count equal to the threshold value, the 0.9-quantile in this case, is disregarded here, as it is of no consequence with respect to the central issue of this analysis, the association of publication month and indicator values. The computed values for selected citation windows are presented in Table 3.4. In that table, the row labeled “actual share of publications $\geq$ threshold” shows the exact percentage of articles that have as much or more citations than the threshold value, which is given one row above it. In the absence of bias, each month’s percentage

should be very close to that overall percentage, which is the expected value, and no pattern in the values should be discernible. For this indicator as well, the month bias can be seen very clearly. For example, in the column for year 1–8 (9-year citation window) in which the expected value is 10.31 %, 12.03 % of January papers reach or exceed the threshold citation count while only 9.46 % of December papers do so. For the 11-year citation window the correlation between publication month and the share of highly cited papers comes to -0.67 . Evidently, citation impact as operationalized by this indicator is also affected by publication month bias, even up to very long citation windows, as is also indicated by the correlations.

Table 3.4: Share of highly cited publications by month for selected cumulative citation windows

		citation window (year)						
		1-4	1-5	1-6	1-8	1-10	1-12	1-14
citation threshold		15	20	24	33	41	49	56
share of publications $\geq$ threshold [%]		10.23	10.06	10.46	10.31	10.33	10.14	10.15
month	1	13.65	12.72	12.78	12.03	11.81	11.35	11.26
	2	11.32	10.67	10.78	10.38	10.28	10.01	9.92
	3	12.22	11.69	11.88	11.37	11.30	10.96	10.89
	4	11.38	10.93	11.20	10.83	10.71	10.39	10.35
	5	10.63	10.31	10.58	10.33	10.26	10.08	10.11
	6	9.85	9.68	10.02	9.91	9.85	9.69	9.72
	7	11.01	10.73	11.20	11.05	11.02	10.82	10.87
	8	8.99	8.93	9.45	9.45	9.53	9.44	9.49
	9	9.61	9.76	10.30	10.27	10.34	10.24	10.24
	10	8.58	8.85	9.54	9.74	9.92	9.87	9.98
	11	8.17	8.47	9.11	9.27	9.55	9.52	9.55
	12	7.89	8.41	9.15	9.46	9.75	9.63	9.72
Pearson r		-0.94	-0.91	-0.87	-0.80	-0.73	-0.67	-0.62
[confidence interval]		[-0.98; -0.79]	[-0.97; -0.70]	[-0.96; -0.58]	[-0.94; -0.42]	[-0.92; -0.26]	[-0.90; -0.16]	[-0.88; -0.07]

3.4.3 The size of the publication month effect in different disciplines and its relation to citation dynamics indicators

The analysis up to this point was concerned with all of articles from the year 2000 as a whole. However, scientific disciplines differ widely in their publication and citation patterns, of which the temporal dynamics patterns are most relevant to this study. Because of these particularities it cannot be readily assumed that the month effect found for the general case is necessarily present or of

substantial size in all scientific disciplines. Hence we now proceed to study the possible field differences of the month effect. As it was shown that the month effect originates in the citations of the first three years, including the publication year, the analysis will be restricted to this three-year citation window, to maximize the bias relative to the total citations. For establishing the incidence of the effect, the correlation coefficient will be used as before and to quantify its relative magnitude the value of the December papers' CPP will be given as a percentage of the January papers' CPP. As for the relationships between the temporal citation dynamics of the disciplines and the month bias, the analysis will take into account the age of the references (cited half-life, i.e. fractional median of the age of the cited literature in years) and the immediacy index as an indicator of the quantity of citations in a short period after publication (average CPP in the publication year) and the ratio of the first year CPP to three-year CPP, here called relative immediacy, as an indicator of the relative concentration of citations in first year, or the skewness towards the beginning of the citation period. The latter indicator was considered because the immediacy index was almost perfectly correlated with 3-year citations. For a disciplinary classification, the OECD Fields of Science classification was used, for which Clarivate Analytics maintains a mapping to its WoS subject categories.

Of the 39 disciplines, the correlations between months and 3-year CPP are significant for all but five disciplines, all of which are social sciences or humanities disciplines with very low citation levels as reflected in 1-year and 5-year average CPP, see Table 3.11 in the Appendix in Supplementary material. The bias is largest in Law, Veterinary science, and Psychology. It is relatively small in Educational sciences, Mechanical engineering, and Industrial Biotechnology. For the calculation of the following correlations of the bias size with the other indicators, the values for those five fields where no bias could be detected were discarded.

The resulting correlation coefficients between the disciplines' relative bias and the other citation indicators are low to intermediate with confidence intervals including zero, see Table 3.5. For this data, the relative magnitude of the month bias is not closely related to any of the citation indicators that were computed: the average CPP after one and three years, the relative immediacy of onset of citation and cited half-life.

Table 3.5: Correlations between month bias size and citation indicators across disciplines (n=34)

	cited half-life	avg. CPP 3y	immediacy index (avg. CPP 1y)	relative immediacy
correlation with relative bias size [confidence interval]	-0.20 [-0.50; 0.15]	0.34 [-0.00; 0.60]	0.31 [-0.04; 0.58]	-0.24 [-0.54; 0.10]

3.4.4 Effect of publication month at the level of individual publications

In the preceding sections publications were aggregated to month cohorts to better illustrate the date dependence of citation counts. This mode of analysis is not particularly useful beyond such demonstration. Regarding the concerns about the validity of bibliometric studies, the influence of publication month on the level of individual papers is more relevant. We turn now to the association between citation counts of individual articles and publication month. Contrary to the averages and ratios used earlier for month-based groups of papers, the distribution of individual article citation counts exhibits severe skewness. Pearson correlations are intended to measure the linear correlations between normally distributed variables. The log-transformation of citation count plus 1 reduces the skewness. Therefore the results in Table 3.6 are given as correlations for both untransformed and log-transformed citation counts for selected citation windows. The calculated associations are small but not zero. The reason for the low values is the very large variability of citation counts due to reasons other than publication month.

Table 3.6: Correlation of publication month and individual article citations for selected citation windows

	citation window and transformation					
	3y cits	log(1 + (3y cits))	5y cits	log(1 + (5y cits))	10y cits	log(1 + (10y cits))
Pearson r, entire data set [confidence interval]	-0.068 [-0.070; -0.066]	-0.097 [-0.099; -0.095]	-0.036 [-0.038; -0.034]	-0.048 [-0.051; -0.046]	-0.015 [-0.017; -0.012]	-0.019 [-0.021; -0.016]

Note:

“3/5/10y cits” is short for 3/5/10 year cumulative citation counts.

3.4.5 Effect of online publication month

Up to this point, the publication month of the print issue has been studied. It was also mentioned that issues may be available a considerable length of time before or after the indicated month. Furthermore, individual articles might be published online, usually in advance, at a date quite different from the publication date of the issue. This raises the question if it might not be better to use the online publication date instead of the issue date to eliminate publication date bias. We now turn to data set B to study this issue.

We use linear regression to assess the unique and combined relationship of print and online publication months on papers' three-year cumulative citation counts. The dependent variable is the natural logarithm of citations after three years plus one, in order to account for the heavy skewness of the citation distribution (variable *cits3y*), following the recommendation in Thelwall (2016). The transformation results in a variable with mean = 1.79, SD = 0.95 and skewness = 0.13.

Print publication month and online publication month are counted as months since January of year 1 CE, so that for example January 2009 has the value 24109. The correlation between the two variables is 0.771 (CI: [0.770; 0.773]). This clearly shows that for papers published in the same month according to the print issue there is considerable variation in the online publication month. Knowing one of the dates only gives a rough indication of the value of the other. For example, for the print issue month publications of January 2009 (N = 22,582), only 4.6 % appeared online in the same month and a further 0.7 % appeared later. Most papers appeared online in the 5 months immediately before January 2009 (79.5 %)

The Pearson correlation of print publication month and citation count on the article level for a three-year citation window for this data set is -0.081 (CI: [-0.084; -0.077]) and the correlation of month and $\log(1 + 3y \text{ cites})$ it is -0.144 (CI: [-0.148; -0.140]), which confirms that the basic month effect is present in this data set as well. The correlation of online publication month and three-year citations is -0.007 (CI: [-0.011; -0.003]) and that of online publication month and log-transformed citation count is -0.040 (CI: [-0.043; -0.036]).

Four regression models for different specifications of the independent variables are used to investigate the individual and combined influence of publication

month in print and online of papers from the same year on citation counts. The three independent variables are *pub.month* (print issue publication month), *epub.month* (online publication month) and *abs.d* (absolute value of the difference between *pub.month* and *epub.month*). The absolute value was taken to linearize the variable for the difference of publication dates as inspection of the scatterplot of the difference of *pub.month* and *epub.month* and the conditional average of *cits3y* showed that the average citation count is highest for papers with no online-to-print delay and that there are decreasing values for papers published both earlier and later online than in print, see Fig. 3.2.

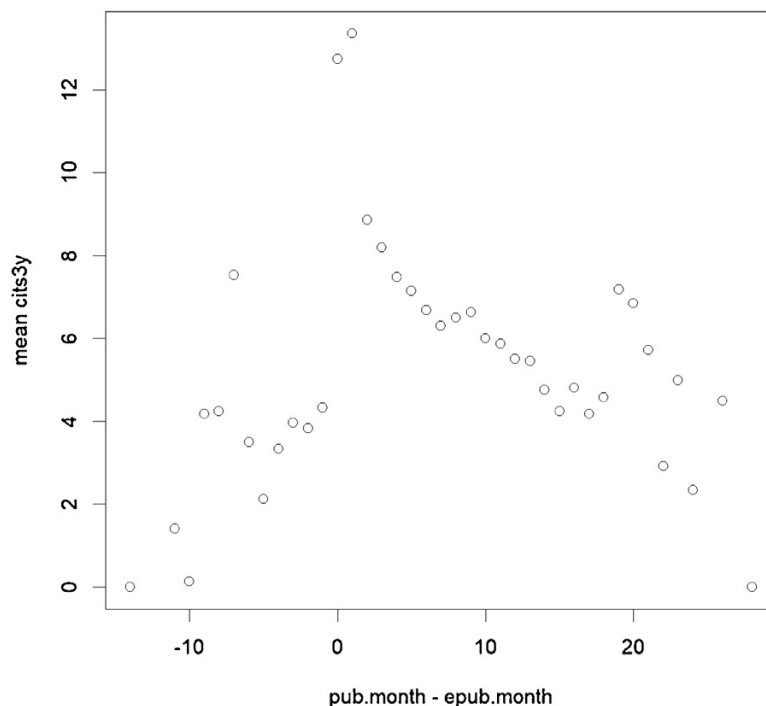


Figure 3.2: Scatterplot of difference in print publication month to online publication month and conditional average of 3-year citations

The results of specifications (1) and (2) in Table 3.7 confirm that both print issue month and online publication month are individually significant predictors of citation counts, being responsible for a 4 % reduction and a 1 % reduction in 3-year citation count on the original scale per increase in one month. The overall predictive power of either variable is low, that of online publication month by itself is negligible.

Table 3.7: Regression of 3-year citations on print and online publication months

	dependent variable: $\log(1 + \text{cits3y})$			
	(1)	(2)	(3)	(4)
pub.month (SE)	-0.040* (0.001)		-0.078* (0.001)	-0.039* (0.001)
epub.month (SE)		-0.009* (0.0004)	0.039* (0.001)	
abs.d (SE)				-0.043* (0.001)
R ²	0.021	0.002	0.034	0.036

Note:

253,292 observations. Constants not shown. * $p < 0.01$.

In model (3) both month variables are entered, which increases their coefficients in absolute terms but changes the sign of *epub.month*. The total explained variation is also greater than the sum of those of models (1) and (2). The positive sign on *epub.month* shows that when print issue month is held constant, those papers with higher online publication month (being published online closer to being published in print) have higher citations on average. This is likely related to the low online-to-print delay in high-impact journals noted above and displayed in Fig. 3.2.

Specification (4) includes the variable *abs.d* in addition to *pub.month*. This leads to an improvement in total predictive power over specification (1) comparable to (3). While their correlation is high, the two publication date variables are complementary to some degree. Online publication month by itself only accounted for very little citation count variation compared to issue month, however this might be different in more recent data. It should be noted that because journal impact covaries with print-to-online publication lag specifications (3) and (4) may also capture citation count differences related to journal impact.

3.4.6 Month level precision citation windows

It has been shown that different month cohorts of publications in the same year have systematically different citation distributions. It was also found that the effect can be detected on the level of individual publications. The consequences for research using regression methods are as simple as including the publication month as a control variable. This approach is not possible in research using other methods and in bibliometrics-informed research assessment. However, instead of correcting or controlling for publication month bias after the data have been

collected, it is possible to collect data in way so that the bias is not introduced in the first place.

One way to prevent introducing the distortion is to use *month-based citation windows*. This approach requires publication month data for both the cited and the citing documents. With the term month-based citation windows we refer to the counting of the number of citations to papers within a specific number of months after publication; the length of this period being equal for all publications, as opposed to the conventional year-based citation windows, thus avoiding the month bias. For this, one does not need to alter the period of investigation; one can look at the same papers of, say, one specific year, as when using the conventional citation windows method. In other words, for the calculation of relative citation indicators using reference sets one is not limited to only publications of the same publication month, which would be another possible method of removing the month bias. While all month-based citation windows have the same duration, they do not cover the same period of the calendar, i.e. the citation window of the later publications months in a year encompasses later citation months as that of the earlier months of the same year.

To verify this proposal, the publication month data was constructed for all documents citing the articles in data set A. We used the method described above to estimate missing values of citing papers' publication months. New citation counts based on month-based citation windows were calculated for 12, 36, 60 and 180 months. The citation windows for each month cohort are in this case of the same length, not of lengths differing up to 11 months. The month-based citation windows are also longer then the corresponding year-based windows. Take for example the 3-year citation window and the 36-month citation window. The three-year window ranges from 36 months (from January, year 1, to December, year 3) to 25 months (from December, year 1, to December, year 3). Of course, all month-based citation windows are 36 months long. Table 3.8 shows the average citation counts for some selected citation windows which may be contrasted with the simple year-based method results which can be calculated from the figures in Table 3.2, Table 3.3. In the former case, there is no correlation between publication month and mean CPP. Fig. 3.3 shows the differences in average citation counts of each month's publications for the two citation window methods in the 3-year case. The proposed method fully corrects the bias implicitly contained in conventional citation windows

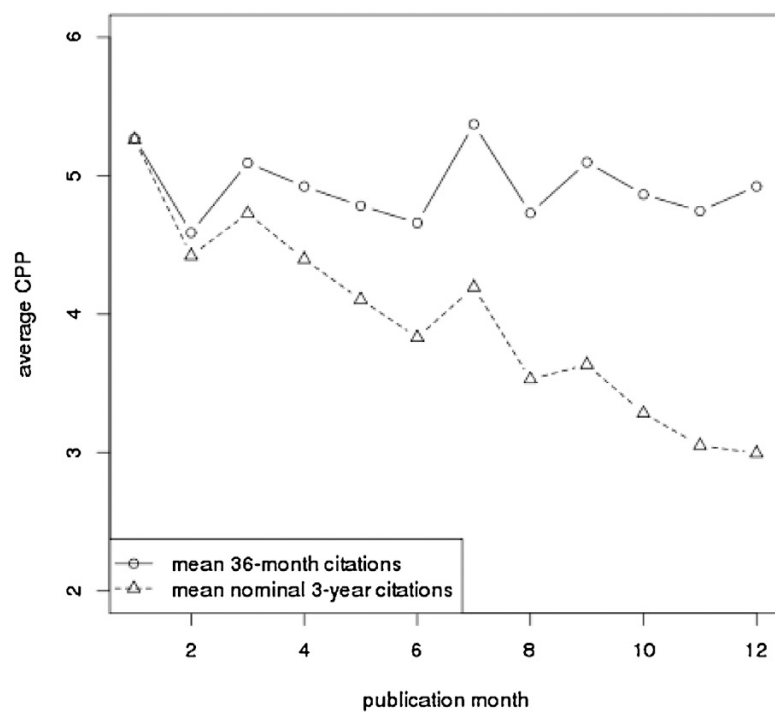


Figure 3.3: Comparison of average citation count of month cohorts using 3-year and 36-month citation windows

Table 3.8: Average CPP, month-based citation windows

		citation window (months)			
		12	36	60	180
month	1	0.87	5.27	9.75	27.69
	2	0.76	4.59	8.55	24.92
	3	0.87	5.09	9.42	26.89
	4	0.84	4.92	9.12	26.04
	5	0.83	4.78	8.83	25.24
	6	0.82	4.66	8.63	24.85
	7	0.98	5.37	9.79	27.08
	8	0.84	4.73	8.70	24.77
	9	0.94	5.10	9.36	26.18
	10	0.89	4.86	9.01	25.39
	11	0.88	4.74	8.71	24.28
	12	0.90	4.92	9.10	25.35
Pearson r		0.50	-0.14	-0.21	-0.51
[confidence interval]		[-0.11; 0.83]	[-0.66; 0.47]	[-0.70; 0.42]	[-0.84; 0.09]

3.4.7 Publication month bias in citations in practice – the influence on institution impact scores

We now turn to an assessment of the practical relevance of the publication month bias. For the purpose of illustration we look at the level of institutions, which are commonly the unit of analysis in bibliometrically informed research evaluation.

We can expect citation based scores to be affected from bias due to the month effect if and in so far as the distribution of publication months of some units deviates substantially from the expectation. As pointed out earlier, under the assumption of random publication months for a set of papers, the expected average month is 6.5. The first point of interest is then to assess if deviations in the average publication month of the publications of specific institutions can be observed. To get an overview, we calculate the average and the standard deviation of the publication month of all publications of each German institution, published in 2000 (data set A). Institutions with 20 or fewer publications are excluded, which leaves 279 institutions. There are 21 institutions in this data set that deviate by one month or more from the expected value in their average

publication month. They are listed in Table 3.9. From those figures, institutions exhibiting a publication output with such a degree of early or late average publication month appear to be only those with a modest amount of output.

Table 3.9: German institutions with relatively low and high average publication months

rank according to average month	average publication month	SD of publication month	publications in 2000	type
1	4.76	3.16	46	company
2	5.00	3.00	39	Leibniz institute
3	5.04	2.74	24	hospital
4	5.10	3.28	23	company
5	5.10	3.28	23	company
6	5.15	3.57	52	Leibniz institute
7	5.25	3.57	26	Fraunhofer institute
8	5.29	3.31	42	Fraunhofer institute
9	5.38	2.87	37	Max Planck institute
10	5.44	3.29	32	company
11	5.48	3.42	23	Fraunhofer institute
270	7.50	3.23	22	Leibniz institute
271	7.50	3.37	30	hospital
272	7.52	3.27	141	Max Planck institute
273	7.52	3.70	23	university
274	7.68	2.88	22	academy of sciences
275	7.72	3.60	50	company (hospitals)
276	8.08	3.34	56	company
277	8.15	4.04	26	hospital
278	8.70	3.84	57	Fraunhofer institute
279	9.94	2.83	31	European research consortium

However, even if a unit's publications are distributed uniformly over the months of a year it does not necessarily mean that its citation scores will not be affected by month bias. This can occur if a unit's high-impact publications in a year are concentrated towards the beginning or towards the end of the year. Compared to other units its scores will appear higher or lower than an unbiased assessment would reveal. Consequently, only considering the average publication months is not enough to rule out distortions due to publication months. Nonetheless, all other things being equal, units exhibiting lower or higher than expected average publication months will have inflated or depressed scores, respectively.

Usually, the research performance of institutions is not assessed across the whole of all science at once but separately by disciplines, in accordance with institutions' disciplinary profile. This reduces the number of publications per insti-

tution that are subjected to citation analysis. To give one example, the most recent national scale assessment exercise of public higher education institutions' research in Australia, ERA 2015, used citation analysis in a number of fields (ARC, 2015). In this case, the minimum number of publications per institution and field was set at 50. Under these conditions, viz. assessments of institutions on the discipline level, can differences in the distributions of publication months between institutions cause significant differences in citations scores? A simple mock evaluation of the German institutions was conducted to uncover the potential distortions that are possible in such a scenario. In this exercise, as we are using data set A, only publications from the year 2000 are included. Each publication and its citations are counted fully for each institution named in the affiliation information. We applied field-normalization at the Disciplines level of OECD Fields of Science classification, using the concordance table of Web of Science categories provided by Clarivate Analytics. Two versions of all German papers' mean field-normalize citation rate were calculated, one in the conventional way, using a 5-year citation window and one with observed and expected citations with month-based citation windows of length 60 months. The results are visualized in Fig. 3.4. In this figure, the lines connect the two data points corresponding to values of the mean field-normalized citation rate using year-based (Y) and month-based citation windows (M) for an institution in a discipline. A flat line indicates no score change (left side panels) or no rank change (right side panels) whereas sloped lines show such changes.

We find that under the condition of a minimum of 50 papers per institution (panels a and b in Fig. 3.4) many institutions' citation scores are corrected, mostly to a small degree. These relatively small changes, however, lead to considerable changes in their rank order (crossed lines in panel b). Increasing the inclusion threshold to 100 papers and thereby selecting the larger institutions with ostensibly less volatile scores (panels c and d) does in fact somewhat dampen the observable score changes. Here, disciplines with fewer than 30 competing institutions have been discarded. Nevertheless, in all disciplines considered a number of rank changes can be found. Many of them are jumps of two or more ranks.

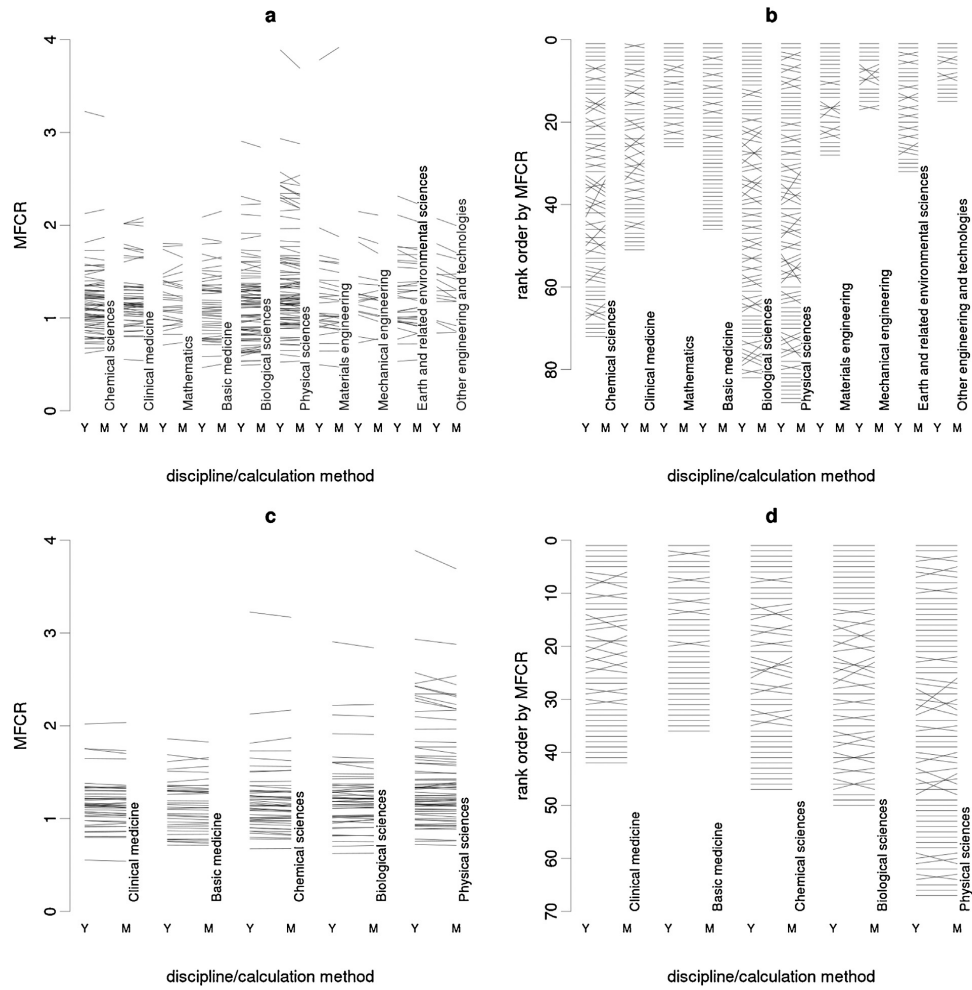


Figure 3.4: German research institutions' citation scores and ranks calculated using conventional and month-based methods.

Notes: Each line connects the positions of the same institution in the year-based and month-based calculations. Crossed lines indicate changed positions. Panels a and b: institutions with more than 50 publications per discipline. Panels c and d: institutions with more than 100 publications per discipline. Y: year-based citation window. M: months-based citation window. MFCR: mean field-normalized citation rate.

3.5 Discussion

3.5.1 Limitations

The present study has some limitations. The study considered only the publication years 2000 and 2009. No temporal dynamics in publication month bias were investigated. The variable of interest was the point of publication of individual articles within a year which was approximated by using the recorded or estimated month of publication of an issue. The publication month was estimated in a portion of the data but the estimation method was shown to work well. Another limitation is that the official month of publication must not necessarily be close to the time when the article could first be read. In fact, individual article dates of online availability would be more accurate but were likely not complete and reliable for the first period studied. For more recent publications the exact date of online publishing would certainly be a better control variable than issue month. Furthermore, the entire subject of publication of works prior to journal publication in different format such as conference proceedings or preprints was beyond the scope of this study.

3.5.2 Discussion of findings

It has been demonstrated that publications of early months have a citation advantage compared to publications of later months of the same publication year. This advantage is of considerable magnitude and cannot be eliminated by removing author self-citations (see Appendix in Supplementary material). The bias is present independent of the particular impact indicator chosen.

Supplementing missing publication month data can be accomplished relatively easily by estimating the publication month from issue sequence information, a method which was shown to have good accuracy. For the studied sample, the effect arises in citations from the first three years after publication but influences metrics using cumulative citation windows for up to 14 or 15 years, for CPP and share of highly-cited papers, respectively. Using longer citation windows can reduce, but not remove, the distortions. These distortions were found for all of science in general and for most disciplines. While the distortions are most readily seen when comparing month cohorts of papers from one year, they can also be detected at the level of citation counts of individual papers. In regression analyses which use citation counts as the dependent variable, print or online

publication month should routinely be included as control variables, in order to account for otherwise unexplained variation, which will improve the estimation of coefficients of variables of substantive interest. Further results clearly indicate that considerable improvements in the accuracy of advanced citation scores are attainable with a minor change in how citation windows are defined, namely by using the month-based citation windows introduced in this paper instead of conventional year-based citation windows. The practical relevance of the findings was demonstrated using a realistic mock evaluation exercise for real research institutions and their real citation scores. The scores exhibited substantial correction effects when recalculated using month level precision citation windows. Month-based citation windows increase the validity of impact indicators which use citation counts by removing systematic error.

Acknowledgements

The author thanks Nees Jan van Eck for providing data set B and for stimulating discussions which greatly improved this study. Part of the analysis was conducted using infrastructure funded through BMBF project 01PQ17001.

Appendix

Author self-citation

It is not immediately clear if author self-citations should be excluded in such an analysis or not. On the one hand, retaining them would lead to some degree of undesirable self-citation contamination in particular in the time immediately following publication because authors have advance knowledge of their own publications that are under review and in press. For the issue studied here, this would indicate that it might be advantageous to exclude self-citations in order to make sure that the association between self-citation and time does not influence the sought after relationship between publication month and citations in general. On the other hand, excluding all self-citations might bias against highly productive or collaborative authors and fields because, if author lists are long, the chances of overlap of author lists of the cited and citing work increase. However, the increase in self-citations is lower than the increase in foreign citations (Glänzel & Thijs, 2004). Using the fuzzy detection method from Donner (2016b) all citations were analyzed and classified into first author/first author self-citations, other author self-citations and non-self-citations.

In order to assess the effects of self-citations on the distribution of citations over months we computed the ratio of the average citations per paper (CPP) of December papers to the average CPP of January papers in the year of publication and the next two years and all three years together, in each case using all citations, only self-citations or removing any self-citations. If there was no month bias at all, the ratio would be 1.0. The results are shown in Table 3.10.

Table 3.10: Ratio of mean CPP of December papers to mean CPP of January papers

set of citations counted	(mean CPP of December papers) / (mean CPP of January papers)			
	year 1	year 2	year 3	years 1 through 3
all citations	0.078	0.470	0.838	0.568
self-citations	0.114	0.704	1.064	0.680
excluding all self-citations	0.058	0.398	0.786	0.534

In the “all citations” row it can be seen that the mean CPP for December papers is just 7.8 % of that of January papers in the first year, rising to 47 % and 83.8 % in the second and third year, respectively. The situation is quite different when

considering self-citations only. By the third year, December papers are self-cited just as much as January papers. This confirms that authors, as citers of their own papers, are comparatively less responsible for any early month bias. However, as the fourth column shows, in the three-year aggregate, there is not a substantial difference between the ratios when calculating with or without author self-citation. The influence of self-citations is likely minor because they account for a small proportion of all citations. As author self-citations only slightly attenuate early months bias in citation counts, they are not excluded in the main analysis of data set A.

Month bias in different disciplines

Table 3.11: Month bias and citation indicators for disciplines in the Fields of Science classification, data set A

discipline	articles	correlation of numerical month and month avg. CPP	month bias (relative bias size Dec : Jan)	cited half-life	immediacy index	avg. CPP after 3 years	relative immediacy
Law	3649	-0.75	34.59	5.48	0.32	1.63	19.50
Veterinary science	9987	-0.86	38.49	7.94	0.16	1.63	10.09
Psychology	21842	-0.89	40.49	8.31	0.26	2.88	8.99
Sociology	11393	-0.94	41.60	7.77	0.14	1.30	11.07
Nano-technology	4048	-0.74	44.00	5.21	0.22	3.11	7.13
Languages and literature	11553	-0.58	46.85	12.54	0.04	0.33	13.71
Health sciences	41550	-0.95	47.42	6.02	0.29	3.29	8.73
Social and economic geography	6433	-0.83	48.16	5.98	0.15	1.46	10.36
Other humanities	2359	-0.04 (n.s.)	48.60	12.14	0.04	0.23	18.05
Clinical medicine	199278	-0.97	49.14	5.85	0.34	4.64	7.34
Materials engineering	47057	-0.94	50.28	7.23	0.16	2.19	7.25
Animal and dairy science	4384	-0.89	50.92	8.29	0.17	1.69	10.20
Basic medicine	98364	-0.94	51.56	5.73	0.43	5.49	7.90
Chemical engineering	12628	-0.96	53.47	7.90	0.15	1.90	7.87
Earth and related environmental sciences	47274	-0.96	54.68	7.81	0.31	3.02	10.17
Other engineering and technologies	34365	-0.93	57.46	7.05	0.19	1.92	9.98
Agriculture, forestry, and fisheries	17603	-0.93	59.75	8.31	0.20	2.03	9.82
Other agricultural sciences	10024	-0.88	60.15	7.72	0.15	2.02	7.29
Biological sciences	159852	-0.96	60.32	5.73	0.59	6.60	8.89
Environmental biotechnology	12593	-0.88	61.15	5.69	0.33	4.40	7.58
Civil engineering	10890	-0.90	61.32	6.42	0.14	1.81	7.93
Electrical engineering, electronic en., information en.	36788	-0.94	61.41	5.59	0.12	1.42	8.40
Political Science	5830	-0.92	61.54	5.22	0.13	1.11	11.88
Computer and information sciences	20774	-0.96	61.79	5.95	0.13	1.45	9.04
Economics and business	20101	-0.84	62.22	7.28	0.14	1.39	10.16
History and archaeology	7253	-0.49 (n.s.)	63.00	14.60	0.07	0.39	18.29
Chemical sciences	111275	-0.96	64.13	6.95	0.32	3.78	8.33
Medical engineering	5818	-0.90	64.36	6.24	0.21	2.93	7.18
Environmental engineering	16161	-0.86	65.39	7.29	0.13	1.66	7.94
Physical sciences	116127	-0.90	66.71	5.93	0.47	4.42	10.59
Mathematics	37594	-0.96	67.94	8.99	0.15	1.59	9.41
Media and communications	1876	-0.78	68.15	5.07	0.24	1.45	16.54
Educational sciences	5495	-0.76	75.71	6.61	0.12	1.19	10.30
Mechanical engineering	30009	-0.88	77.52	8.04	0.14	1.46	9.89
Other natural sciences	8761	-0.67	81.85	4.48	1.72	18.04	9.53
Industrial Biotechnology	1179	-0.71	82.98	6.77	0.25	4.04	6.10
Other social sciences	897	-0.18 (n.s.)	99.52	10.43	0.09	0.77	11.94
Philosophy, ethics and religion	5268	-0.16 (n.s.)	109.86	11.11	0.08	0.49	16.52
Art (arts, history of arts, performing arts, music)	8099	-0.09 (n.s.)	142.13	12.97	0.03	0.11	23.46

Note:

Disciplines are sorted by the relative size of the month bias, in descending order of the strength of the bias. Correlations coefficients in column 3 are statistically significant at the 5 % level unless noted.

4 Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems

This chapter was first published under the same title in *Quantitative Science Studies*, vol. 1, issue 1, pp. 150-170, at https://doi.org/10.1162/qss_a_00013, on 01 February 2020. The authors are Paul Donner, Christine Rimmert (Bielefeld University, Germany), and Nees Jan van Eck (Centre for Science and Technology Studies, Leiden University, Netherlands).

Abstract

The present study is an evaluation of three frequently used institution name disambiguation systems. The Web of Science normalized institution names and Organization Enhanced system and the Scopus Affiliation ID system are tested against a complete, independent institution disambiguation system for a sample of German public sector research organizations. The independent system is used as the gold standard in the evaluations that we perform. We study the coverage of the disambiguation systems and, in particular, the differences in a number of commonly used bibliometric indicators. The key finding is that for the sample institutions, the studied systems provide bibliometric indicator values that have only a limited accuracy. Our conclusion is that for any use with policy implications, additional data cleaning for disambiguating affiliation data is recommended.

Keywords

research evaluation, institution name disambiguation, bibliometric data, data quality, Web of Science, Scopus, validation study

4.1 Introduction

Scientometric studies at the level of research institutions face the challenge of the correct attribution of publications to institutions. This task, here referred to as *institution name disambiguation*, comprises systematically standardizing the heterogeneous address data of the author-provided affiliation information present in publications and recorded in bibliographic databases. At present,

institutional affiliation information in academic publications is not standardized and unique identifiers for research institutions have not yet been adopted. Therefore, in order to generate valid primary data on publications for studies at the meso level, the assignment of address strings to known real institutional entities is crucial. Institution name disambiguation belongs to a class of problems known as named entity normalization, in which variant forms have to be matched to the correct preferred form. Another prominent member of this class is author name disambiguation. Disambiguated affiliation information can contribute to the performance of author name disambiguation systems that employ affiliations as background information.²⁰

In the recent past, a nearly complete institutional disambiguation for German research institutions was developed and implemented at the Institute for Interdisciplinary Studies of Science at Bielefeld University, as a major component of a national bibliometric data infrastructure for research and monitoring (Rimmert et al., 2017). The system has been tested and improved over a number of years and is now in production use. We are therefore in a position to study the degree to which the use of a sophisticated disambiguation system with near-complete national-scale coverage leads to different bibliometric indicator values compared to a situation in which no such system is available and simpler alternatives to the attribution problem have to be used. We consider here (a) the case where a simple unification strategy using ad hoc lexical searches in the address data fields of a bibliographic database is conducted in order to collect publications of the target institutions (based on vendor preprocessed affiliation data in Web of Science [WoS]); and (b) the use of bibliographic database vendors' own institution disambiguation systems (in both WoS and Scopus). We believe that these two situations are common in practice outside of specialized research or evaluation units with access to the raw data of bibliographic databases. The performance and implications of these approaches are therefore relevant and of wide interest to the bibliometrics and research evaluation communities. Prominent examples with profound science-political consequences of the use of bibliometric data of institutions derived from WoS or Scopus are the international university rankings, which generate much public attention and elicit considerable debate.

The remainder of the article is structured as follows. We begin by providing

²⁰Likewise, disambiguated author information could potentially be used as additional input information for institutional disambiguation. However, we are not aware of any literature on this approach.

an overview of the prior work on institutional disambiguation. Next, we briefly outline the institution name disambiguation systems that we study and describe the publication data and institution samples that we use, as well as the comparison scenarios, the bibliometric indicators that are calculated for the institutions, and our approach of assessing the differences in indicator values. In the next section we present the results of our comparisons. In particular, we assess the distributions of errors in indicator values over institutions arising when applying alternative disambiguation systems in contrast to the reference values obtained from the presented disambiguation system for German research institutions, which can be assumed to be complete and nearly error free for the data. The results and their implications are summarized in the Discussion section.

4.2 Related work

Unification of author affiliation information and the allocation to clearly identified research institutions has been recognized as a challenging task in the bibliometric research community and beyond. Accurate disambiguation of heterogeneous affiliation data is crucial for institution-level scientometric research and bibliometric evaluation. Disambiguation systems connecting heterogeneous author affiliations to known research institutions have been constructed in several projects – usually for project-specific purposes and not to be made available publicly. They may be roughly divided into rule-based and machine learning approaches. However, this division is not a strict one, as approaches often use combinations of methods (e.g., rules and some manual work are used in addition to a machine learning approach to improve precision, especially for problematic cases).

4.2.1 Rule-based approaches

A substantial amount of work on this topic has been done at the Centre for Science and Technology Studies (CWTS) at Leiden University. For the case of universities, this began with De Bruin & Moed (1990). They performed a unification of about 85,000 affiliations from 75 journals (data from SCISEARCH) using the first part of the addresses. Using structural information and reference books, they assigned units on lower hierarchical levels (e.g., departments), appearing in the first part of addresses, to the corresponding main organization.

They found that many problems remained, and to solve these, they used external information from encyclopedias, university handbooks, specialists, and staff lists of universities. This is a time-consuming method, and they only did this for selected countries (in particular the Netherlands). In a follow-up study, Moed et al. (1995) reported on a bibliometric database constructed from all articles published by authors from the Netherlands using data from the Science Citation Index. To store unified affiliations, they improved their earlier procedure for Dutch addresses by, among other things, adding a classification of institutions to research sectors: that is, types of organizations such as universities, hospitals, and firms. They noted problematic affiliations that could not be handled correctly by their procedure. CWTS continued to maintain and improve its disambiguation system, in particular for its university ranking, for which all name variants that occur at least five times in the WoS are cleaned (Waltman et al., 2012). This system pays special attention to the way publications by academic hospitals are assigned to universities (Reyes-Elizondo et al., 2016).

The Swedish Research Council performed affiliation disambiguation for its bibliometric database, which was also constructed based on WoS data (Kronman et al., 2010; Swedish Research Council, 2017). They used a deterministic approach based on a catalog of string rules, mapping address strings to 600 known Swedish research organizations. Organizations were also classified by the research sector. Their procedure was able to assign over 99 % of Swedish address strings. A single address may be matched to more than one organization in the case of affiliations containing information on more than one organization, usually indicating collaborations.

4.2.2 Machine learning approaches

French et al. (2000) described a number of institutional disambiguation experiments with different address string distance metrics and a one-pass heuristic clustering procedure. The clearly stated goal was not a complete automatic disambiguation, but rather the reduction of manual reviewing of the most difficult cases. Among other things, they introduced a new, domain-specific affiliation comparison function, based on normalized and sorted words, minimizing edit distances between aligned words across possible permutations. Jonnalagadda & Topham (2010) reported on their disambiguation of institution names extracted from PubMed data. The presented approach utilized agglomerative clustering,

for which the entity similarity is computed with an edit distance, building on the work of French et al. (2000). In particular, their approach was a hybrid of a sequence alignment measure over word sequences (Smith–Waterman algorithm) and the Levenshtein distance between single words. Furthermore, similar clusters were merged. They reported sample precision values of 99.5 % (4,135 affiliation strings related to “Antiangiogenesis,” only US organizations) and 97.9 % (1,000 affiliation strings related to “Diabetes,” organizations from any country) for organization normalization. Although these values are high, it is not possible to extrapolate them to less restricted data sets. Galvez and Moya-Anegón (2007; 2006) reported on a new approach using finite-state graphs, developed with WoS data and also tested on data from Inspec, Medline, and CAB Abstracts. Although this is a promising approach, the authors outlined the limits of automatic classifications for problematic affiliations, which requires expert knowledge to classify. Jiang et al. (2011) discussed an experimental approach of agglomerative clustering of affiliations using string compression distance. Their evaluation of the method is questionable, as they use the publication affiliations of mostly students and staff from a single university. Their affiliation string pool is, therefore, dominated by name variants of that university, although the remainder are affiliations of coauthors. They extracted a reference corpus of 217 “affiliations” (variants) of 105 “categories” (true organizations). In any case, their clustering quality results are not encouraging. This also holds true for the application of supervised and semisupervised machine learning methods, tested by Cuxac et al. (2013) on French CNRS addresses. S. Huang et al. (2014) proposed an algorithm using author information to classify affiliations that received high precision values but low recall.

4.3 Systems, data, and methods

4.3.1 Institution disambiguation systems

In this section we summarize the disambiguation system that was developed for German institutions. For a full description of the system we refer readers to Rimmert et al. (2017). The system, which we call the *KB system*,²¹ is comprised of (a) a set of known and uniquely identified German research institutions, (b) a mapping of institutions to affiliation records identified as belonging to each

²¹The system was created in a project called *Kompetenzzentrum Bibliometrie* (Competence Centre for Bibliometrics).

institution from the two data sources WoS and Scopus, (c) a hierarchical classification of the institutions into sectors, and (d) a change history of the institutions which record the splitting and merging and incorporation of institutions and sector changes. The KB system is thus built on the affiliation data provided in WoS and Scopus, respectively, and belongs to the category of rule-based systems. The tracking of structural changes affords the necessary flexibility in handling such changes required by different project contexts. In the KB system, two different analytical views are implemented (item (d) above). With Mode S (for synchronic allocation), we can perform analyses that take into account the institutional structures as they were at the time of publication for each paper. Institutions that have later come to be related to another institution through structural changes, such as through mergers or splits, are treated as different entities. On the other hand, with Mode A (asynchronic, current perspective), we can analyze the publication records of institutions as they are constituted at present; that is, including publications of predecessor units. The mapping of institutions to affiliation records (item (b) above) is a deterministic, rule-based classification. The core of the institutional coding procedure is a mapping of author addresses to the corresponding uniquely identified research institutions and their subdivisions, using a large library of regular expressions. This library currently contains some 45,000 expressions and is continuously being expanded and improved.

The sector classification (item (c) above) contains the classes of higher education sectors (universities and universities of applied sciences), four major nonuniversity research organizations (Fraunhofer-Gesellschaft [FHG], Helmholtz Association [HGF], Leibniz Association [WGL], and the Max Planck Society [MPG]), private companies, registered associations, government laboratories, and academies of science. For the sector information, structural changes over time and multiple assignments of research institutions to these sectors are also available.

The version of the KB system used for this study contained 2,097 institutions, which also included placeholder records for unidentified institutions for which only the sector could be determined. An evaluation of the KB disambiguation system was conducted prior to the main study. We provide a detailed overview of the system evaluation in Appendix A for German research institutions. We conclude that, based on the good results of this evaluation, the KB system is a valid, gold standard benchmark for German institutional affiliation disambigua-

tion data. This is not to say, however, that the KB system or its rule-based approach are superior in general. In fact, its scope is limited to a single country and it would be difficult to extend the method to global scope because of the large effort and unreasonable expense required.

We deliberately do not attempt to describe the workings of the proprietary institution disambiguation systems of WoS and Scopus and regard them as black boxes, of which we only analyze the results. The reason for this is that both systems are not documented in any detail by the providers. What we can gather from the information of the platforms is that WoS Organizations Enhanced (OE) is based on lists of variant names mapped to preferred names.²² WoS OE can therefore be seen as a rule-based system. Regarding the Scopus Affiliation Identifiers (AFIDs), the documentation merely informs us that “the Affiliation Identifier distinguishes between affiliations by assigning each affiliation in Scopus a unique number, and grouping together all of the documents affiliated with an organization.”²³ No information is given about how the system works.

4.3.2 Data

The data used in the analyses are derived from the licensed source of WoS²⁴ and Scopus, obtained in spring 2017. The data were loaded into in-house relational databases, cleaned, and enhanced at the Competence Centre for Bibliometrics for Germany. The most important enhancement is the disambiguation of German author addresses to known German research institutions. This process is conducted separately for each data source using the KB disambiguation system described in the previous subsection. The units of the analysis for this study are German academic institutions, in particular universities, universities of applied sciences, and nonuniversity research institutes. Publications are restricted to articles and reviews published between 1995 and 2017. To be included, an institution needed to have at least 50 such publications associated with it according to the KB disambiguation of the WoS data. These restrictions resulted in a study sample of 445 institutions. The same institutions are used to investigate

²²<https://clarivate.libguides.com/woscc/institution>; <http://ips.clarivate.com/m/pdfs/UnifyingNameVariants.pdf>, accessed April 13, 2018.

²³https://service.elsevier.com/app/answers/detail/a_id/11215/supporthub/scopus/related/1, accessed April 13, 2018.

²⁴The WoS data used in this study include data from the Science Citation Index Expanded, the Social Sciences Citation Index, the Arts & Humanities Citation Index, and the Conference Proceedings Citation Index.

both WoS and Scopus.

4.3.2.1 Scopus AFID For the Scopus data, we compare the KB system-derived reference data to sets of publications that have one or more common assigned AFIDs (affiliation identifiers), as provided by Elsevier. Some preprocessing steps to align the Scopus and KB disambiguation systems were performed in order to make them comparable, as they are conceptually and structurally somewhat different. To match AFIDs to the KB system IDs, the AFID for each institution in our sample was obtained by searching Scopus’s online platform. It is not clear whether and how exactly the definition of an institution in Scopus differs from the one the KB disambiguation is based on. One difference that we have noticed is that the AFID system typically has separate IDs for university hospitals and the universities they belong to, which is not the case in the KB system. We have therefore merged those AFIDs to create more comparable and consistent publication record sets. Furthermore, in some cases more than one AFID for the same institution exists in Scopus, for instance, for multiple branch locations. If these are logically linked in the hierarchical relations in the Scopus system, we also merged these linked AFIDs. If not, we took only the most commonly used AFID per institution.

We found that in the AFID system, publications with affiliations referring to predecessor units are grouped with their current unit. Based on this observation, we compare the AFID results with those from the KB system’s Mode A.

4.3.2.2 Web of Science (WoS) Organization Enhanced The WoS OE system does not have unit identifiers but Preferred Names, which are additionally assigned as institution names to affiliations considered belonging to one real institution. In order to identify the WoS Preferred Name for the institutions in our set, we started by identifying all the Preferred Names of records with German addresses that occur more than 20 times. From this list, we chose the Preferred Name matching the target institution and otherwise excluded the institution from this part of the study. In fact, for our sample set, it was not possible to retrieve the corresponding publications on the main institutional level in a majority of cases. Although many universities are recorded in OE, the institutions of FHG, HGF, WGL, and MPG are almost all grouped such that only all publications of each of the respective head organizations can be found,

but rarely those of their member institutes.²⁵

Similar to AFID, also in the WoS OE system, predecessor units are grouped under the Preferred Name of the current institution. In consequence, we also compare the WoS OE system with Mode A of the KB system.

4.3.2.3 WoS institution name search As well as the comparison of WoS OE data with the KB disambiguation, we also investigated the performance of a lexical search using the institution name in the WoS affiliation data. As pointed out above, the coverage of institutions in the WoS OE system is far from complete (since only head organizations are covered, not their member institutes), which supports the notion that such an alternative approach might often be required in practice. The institution name search method makes use of WoS disambiguation efforts, because institution names extracted from affiliation information in papers are not indexed identically to how they are given in the original publication but are normalized. Because the affiliations in Scopus are not transformed or normalized, we do not apply a similar search strategy to Scopus data. In fact, it is not possible to conduct comparable searches across these two databases because WoS only contains normalized address strings, while Scopus only contains the original address strings.

In this scenario, we model a hypothetical user who has a list of the names of the German research institutions available, which is used as a basis for generating search terms. We also assume that the user is familiar with searching in WoS data to a sufficient degree. This scenario further requires a definition of the name list, the search terms, and the search parameters.

In order to generate a plausible name list, we begin by using the KB institutional disambiguation results to find the most common normalized name in the WoS data for each real institution in our initial set, because in principle there should be only one normalized name for each institution. We manually assess the lists side by side with the real names and discard any WoS name that cannot be deduced from the name list, using instead the next most common name variant iteratively until all WoS normalized names are mapped to KB system IDs based on the names in the two systems. This relates to our decision of going beyond a completely naïve and automatic procedure and including a realistic

²⁵After reaching out to Clarivate Analytics for comments, we were informed that the decision to include the MPG as a whole but not its member institutes was based on a request by MPG. The same approach was then also taken for the other nonuniversity research organizations.

degree of user common sense and domain familiarity. We use the search term list thus obtained as retrieval inputs, while also ignoring capitalization and allowing truncation at the end of the term, and searched the full address information field. This came reasonably close to an informed, but nonspecialized, search for an institution on the online platform of WoS. It is general in the sense that all institutions are treated in the same way and no special knowledge of affiliation idiosyncrasies is included. It is limited in the sense that we only consider one name variant per institution.

Because we directly use the normalized affiliation data as it is indexed in WoS, it is clear that we use the normalized versions of the institution names at the time of publication. Thus, we use Mode S of the KB system for comparison.

4.3.3 Methods

To assess the performance of the studied systems in terms of being able to identify the correct publications of the research institutions we use the information retrieval measures of precision and recall. For this task, precision is calculated as the share of correctly retrieved publications among the total number of retrieved publications. Recall is the share of correctly retrieved publications among all relevant publications. The correct publications of an institution are those identified by the KB system.

In order to quantify the effect of the application of a specific institutional disambiguation on scores of bibliometric indicators, we calculated the indicator values based on the publications of each institution as retrieved by the KB system – considered a validated gold standard for the selected institutions in this study – and for each of the three alternative systems. The differences of indicator values are calculated, and the arising error distributions are displayed.

A number of commonly utilized bibliometric indicators are included in this study. We consider the three domains of publication output, collaboration, and citation impact. For the latter two domains we have selected indicators that are size dependent (absolute numbers) as well as size-independent indicators (ratios or averages). The citation indicators are all calculated for 5-year citation windows which include the year of publication. The indicators are summarized in Table 4.1. It is clear that the size-dependent indicator values are directly related to the number of correctly identified publications. However, it might be hypothesized

that the values of size-independent indicators are less affected when only a part of the correct publication set is used as their input, because errors may cancel each other out.

Table 4.1: Overview of selected bibliometric indicators

Domain	Indicator	Remarks
Publication output	P: Number of publications (full count)	The number of publications of an institution
	P(collab): Number of collaborative publications	The number of an institution's publications that have been coauthored with one or more other institutions
	PP(collab): Proportion of collaborative publications	The proportion of an institution's publications that have been coauthored with one or more other institutions
	P(int collab): Number of international collaborative publications	The number of an institution's publications that have been coauthored by two or more countries
	PP(int collab): Proportion of international collaborative publications	The proportion of an institution's publications that have been coauthored by two or more countries
	TCS: Total citation score	The total number of citations of the publications of an institution
Collaboration	MCS: Mean citation score	The average number of citations of the publications of an institution
	TNCS: Total normalized citation score	The total number of citations of the publications of an institution, normalized for field (WoS: Subject Category; Scopus: ASJC) and publication year
	MNCS: Mean normalized citation score	The average number of citations of the publications of an institution, normalized for field (WoS: Subject Category; Scopus: ASJC) and publication year
	P(top 10%): Number of highly cited publications	The number of an institution's publications that, compared with other publications in the same field and the same year, belong to the top 10 % most frequently cited
	PP(top 10%): Share of highly cited publications	The proportion of an institution's publications that, compared with other publications in the same field and the same year, belong to the top 10% most frequently cited
Citation impact		

We compare two vendor-provided disambiguation system results and one search-based result with the KB system's results, which we take as the correct result providing reference values. We divide the system evaluation into two parts. First, for each institution in the evaluation set, we would like to find all its publications, without retrieving any publications it was not involved in. This is a basic information retrieval task, which can be measured with precision and recall. We also use retrieval performance, including the absolute number of retrieved institutions in the evaluation set, to analyze the coverage of the systems with respect to our sample of 445 institutions. The second component of the evaluation concerns the bibliometric indicator scores calculated from the

retrieved institution publication sets. In general, the numerical discrepancy between the indicator values, using the KB disambiguation (reference values) and the other methods, will be expressed as relative deviation in percent, calculated as

$$\text{deviation} = \frac{(\text{observed system score} - \text{KB system reference score})}{\text{KB system reference score} \times 100}.$$

The deviation has a lower bound at -100% and is unbounded in the positive direction. For example, let the reference MCS of a unit be 5.5 (calculated based on the KB disambiguated data), and the focal value obtained from a simple institution search in WoS be 4.2. Then the deviation as defined above is $(4.2 - 5.5)/5.5 \times 100 = -23.6\%$. In this case, the correct result would be underestimated by 23.6 %.

For each indicator, the computed deviations for each institution are collected. Our main measure of accuracy is the percentage of values within a range of $\pm 5\%$ of the reference score.

4.4 Results

An overview of the coverage of German institutions in the WoS and Scopus institution disambiguation systems and the lexical search method in WoS is provided in Table 4.2. We are able to find only 91 of our 445 (20 %) evaluation sample institutions in the OE system. The coverage of OE is the lowest among the systems considered. To a significant extent, this is a consequence of the choice not to include the member institutes of nonuniversity research organizations in WoS OE. The set of covered institutions in WoS OE is comprised mostly of universities. However, also for the universities, in particular for the universities of applied sciences, a significant number of institutions are not covered in WoS OE. Using the search strategy, we can find one normalized form for each institution, achieving complete coverage of the institutions. The Scopus AFID system covers 376 (85 %) of the institutions with no conspicuous differences between sectors.

Table 4.2: Coverage of sample institutions by the studied disambiguation systems

Sector	Number of institutions	Covered in WoS OE	Covered in WoS search	Covered in Scopus AFID
Fraunhofer-Gesellschaft	62	1	62	57
Helmholtz Association	23	9	23	16
Max Planck Society	86	0	86	78
Universities	107	66	107	96
Leibniz Association	83	8	83	56
Universities of applied sciences	90	9	90	75
Total	445	91	445	376

Note:

The figure in the row “Total” may differ from the sum of the above cells because some institutions are assigned to more than one sector.

4.4.1 WoS Organizations Enhanced

We present the institution-level figures for precision and recall for WoS OE in Table 4.3 and Figure 4.1. All results should be interpreted with due caution because of the OE system’s limited coverage of the selected institutions. The precision of WoS OE for these institutional publication sets is 0.95, on average, across institutions, weighted by publication numbers. Hence, typically about 5 % of the returned publications in a result set of a specific preferred name will be false positives. The weighted mean of recall across institutions is 0.93, meaning that the result sets do not include about 7 % of relevant publications, on average. The contrast between unweighted (0.87) and weighted mean for the recall shows that the results for larger institutions (in terms of number of publication) are better than for smaller institutions. We found poor results for recall for the four institutions presented in Table 4.4.

Table 4.3: Summary statistics of the distributions of precision and recall of retrieved publications per institution for WoS OE ($n = 91$)

	Minimum	Median	Arithmetic mean	Weighted mean	Maximum	Standard deviation
Precision	0.41	0.99	0.96	0.95	1.00	0.11
Recall	0.07	0.94	0.87	0.93	0.99	0.18

Table 4.4: Institutions with low recall for WoS OE

Institution Name (KB system)	WoS OE Preferred Name	Recall
Helmholtz-Zentrum für Infektionsforschung	Helmholtz Center Infectious Research	0.07
Hochschule Fresenius	Fresenius University of Applied Sciences	0.15
Fraunhofer-Institut für Optonik, Systemtechnik und Bildauswertung IOSB	Fraunhofer Optonics, System Technologies & Image Exploitation Ettlingen	0.16
Leibniz-Institut für Festkörper- und Werkstoffforschung Dresden	Leibniz Institute Solid State & Materials Research Dresden	0.18

We now turn to the results of the comparison of the scores of bibliometric indicators between the WoS OE and the KB system. The results are presented

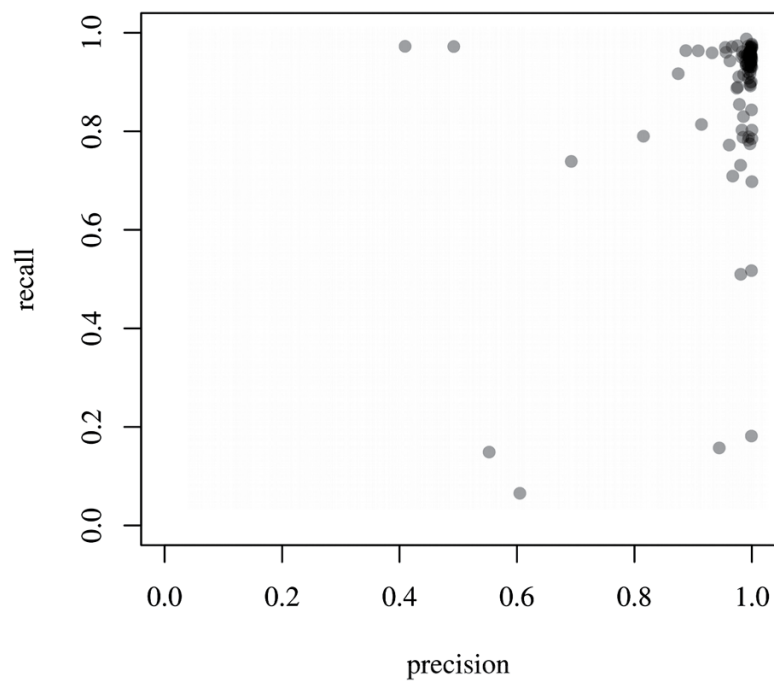


Figure 4.1: Precision and recall for WoS OE ($n = 91$).

in Table 4.5, in the form of summaries of the deviation score distributions, visualized in Figure 4.2. It can be seen that absolute indicator scores (number of publications, collaborative publications, and citations) are less often within the range of nearly correct values ($\pm 5\%$) than relative indicator scores.

Table 4.5: Deviation of indicator scores of WoS OE from KB system ($n = 91$)

Indicator	Percent deviation within $\pm 5\%$	Median absolute deviation (%)
P	40.7	4.50
P(collab)	38.5	4.30
PP(collab)	93.4	0.60
P(int collab)	37.4	3.40
PP(int collab)	86.8	1.30
TCS	52.6	4.00
MCS	79.1	1.30
TNCS	40.7	3.70
MNCS	85.7	1.30
P(top 10%)	48.4	3.70
PP(top 10%)	100.0	0.04

4.4.2 WoS Institution Name Search

In this section, we compare the results of the WoS institution name search with those of the KB system. Note that the search makes use of the institution name normalization of WoS, and we have deliberately searched for the single most common WoS normalized name per institution, as mentioned above. Using this search method, we obtain vastly more institution publication sets than using WoS OE; in fact, full coverage of all sample institutions is achieved (see Table 4.2). The summary of the distributions of precision and recall is given in Table 4.6 and the values are displayed in Figure 4.3. We obtain rather poor results for the average precision of 0.61 when weighting institutions by the number of publications, and 0.67 as the arithmetic mean. Publication sets for this method will often contain many publications incorrectly assigned to the institutions in question. Recall is at 0.74 weighted mean and 0.55 arithmetic mean, which means that the publication lists returned by these queries will commonly be incomplete, but less so for the larger institutions. Tables 4.7 and 4.8 provide the five institutions with the lowest recall and precision scores.

These results for recall suggest that the normalization procedure of WoS is often unable to group most of the relevant institution name variants under one

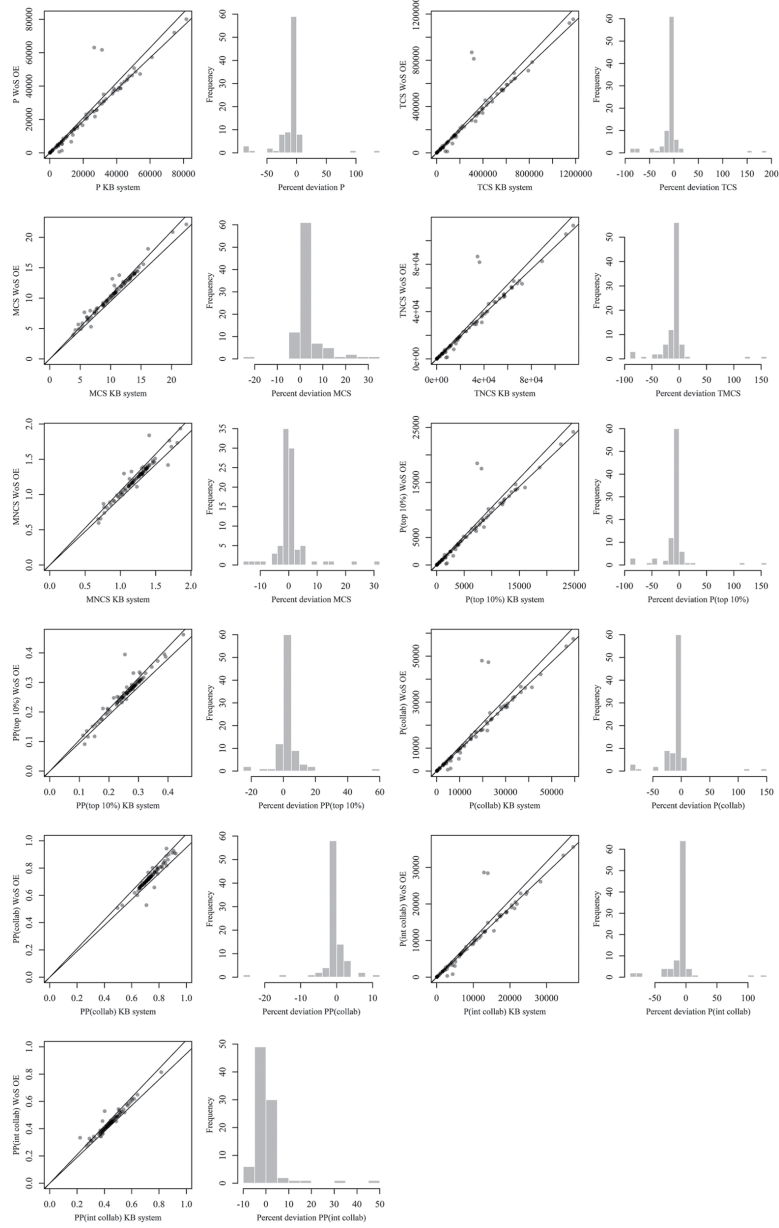


Figure 4.2: Distributions of indicator score deviations of WoS OE from KB system in Mode A. Diagonal lines indicate the $\pm 5\%$ error margin for indicator values.

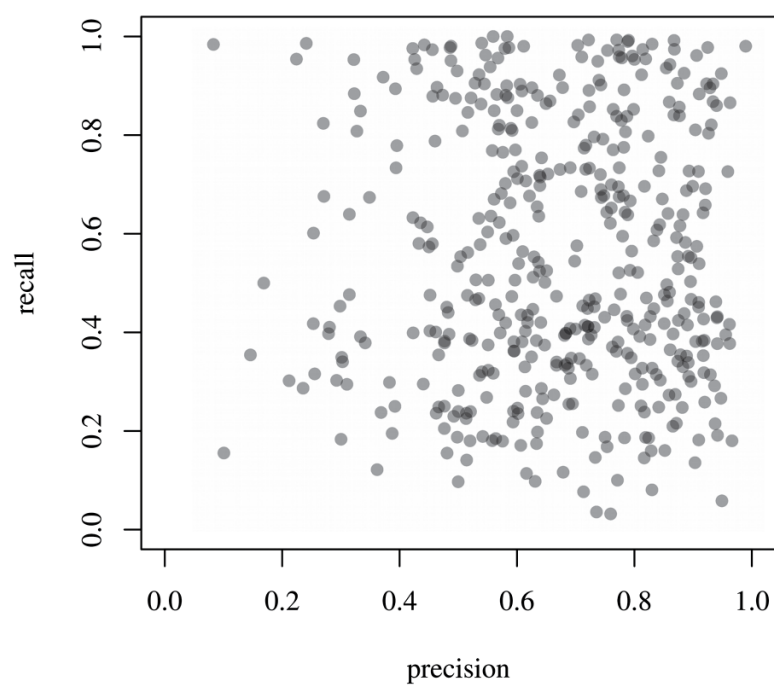


Figure 4.3: Precision and recall for WoS institution name search ($n = 445$).

Table 4.6: Summary statistics of distributions of precision and recall of retrieved publications per institution for WoS institution name search ($n = 445$)

	Minimum	Median	Arithmetic mean	Weighted mean	Maximum	Standard deviation
Precision	0.08	0.69	0.67	0.61	0.99	0.19
Recall	0.03	0.51	0.55	0.74	1.00	0.26

Table 4.7: Institutions with low recall for WoS institution name search

Institution Name (KB system)	WoS OE Preferred Name	Recall
Max-Planck-Forschungsgruppe Optik, Information und Photonik	INST OPT INFORMAT & PHOTON	0.03
Leibniz-Institut für Katalyse e. V. an der Universität Rostock (LIKAT)	LEIBNIZ INST KATALYSE EV	0.04
Friedrich-Miescher-Laboratorium für biologische Arbeitsgruppen in der Max-Planck-Gesellschaft	FRIEDRICH MIESCHER LAB	0.06
Berliner Elektronenspeicherring - Gesellschaft für Synchrotronstrahlung mbH	BERLINER ELEKTRONENSPEICHERRING GESELL SYNCHROTRO	0.08
Hochschule für angewandte Wissenschaft und Kunst, Fachhochschule Hildesheim/Holzminde/Göttingen	HAWK UNIV APPL SCI & ARTS	0.08

Table 4.8: Institutions with low precision for WoS institution name search

Institution Name (KB system)	Most common WoS normalized name	Precision
Max-Planck-Institut für Biologie	MAX PLANCK INST BIOL	0.08
Max-Planck-Institut für Biophysik	MAX PLANCK INST BIOPHYS	0.10
Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB	FRAUNHOFER IOSB	0.15
ESCP Europe Wirtschaftshochschule Berlin e.V.	ESCP EUROPE	0.17
Forschungszentrum für Marine Geowissenschaften GEOMAR	GEOMAR	0.21

normalized form.

The results of the comparison of the bibliometric indicator scores between the WoS institution name search approach and the KB system for Mode S are provided in Table 4.9 and the deviation distributions are displayed in Figure 4.4. The shares of institutions for which the scores obtained with the institution name search approach are within $\pm 5\%$ of the reference score are low, especially for the absolute indicators. Dispersion of the deviations is high. Moreover, the ratio- and mean-based citation scores are comparatively less inaccurate. Evidently, the incomplete publication list result sets of this method lead to substantially inaccurate scores for all indicators.

4.4.3 Scopus AFID

The results for precision and recall of the Scopus AFID system, under the Mode A condition, are summarized in Table 4.10 and displayed in Figure 4.5. Precision is quite high, but, in contrast, recall is more moderate. Again, we find that the weighted mean precision and recall are slightly greater than the unweighted ones, suggesting that disambiguation quality is typically a little better for larger institutions. We also note that the coverage of our selected benchmarking in-

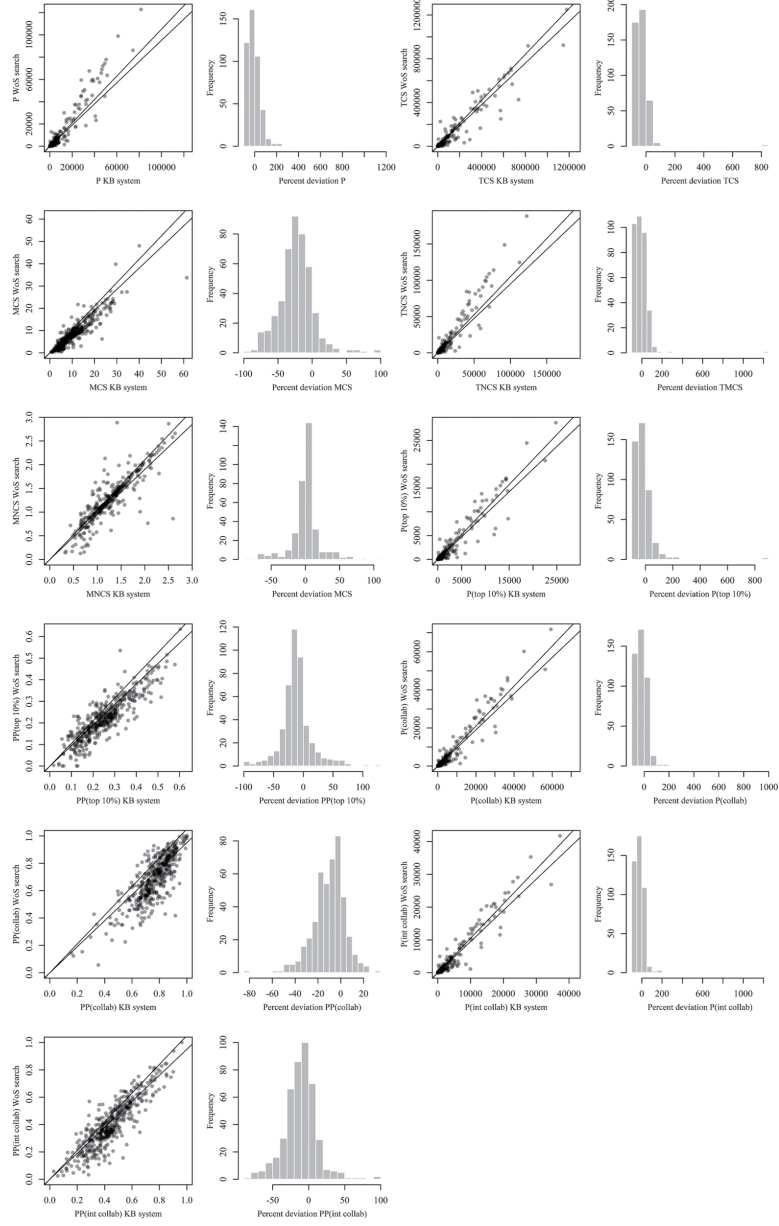


Figure 4.4: Distributions of indicator score deviations of WoS institution name search from KB system in Mode S. Diagonal lines indicate the $\pm 5\%$ error margin for indicator values.

Table 4.9: Deviation of indicator scores of WoS institution name search from KB system ($n = 445$)

Indicator	Percent deviation within ± 5 %	Median absolute deviation [%]
P	6.7	52.5
P(collab)	7.2	45.4
PP(collab)	29.0	12.5
P(int collab)	10.1	45.2
PP(int collab)	20.5	17.9
TCS	10.8	39.6
MCS	9.9	18.8
TNCS	5.7	55.5
MNCS	52.0	6.6
P(top 10%)	7.6	46.6
PP(top 10%)	13.7	14.7

stitutions for the AFID system is 376 out of 445 (i.e., 85 %) and therefore far from complete. Unlike the WoS OE system, the Scopus AFID system is not largely concentrated on universities (Table 4.2). Table 4.11 provides the five institutions with the lowest recall scores for the Scopus AFID system.

Table 4.10: Summary statistics of the distributions of precision and recall of retrieved publications per institution for Scopus AFID ($n = 376$)

	Minimum	Median	Arithmetic mean	Weighted mean	Maximum	Standard deviation
Precision	0.17	0.96	0.92	0.96	1.00	0.1
Recall	0.03	0.90	0.82	0.86	0.99	0.2

Table 4.11: Institutions with low recall for Scopus AFID

Institution Name (KB system)	AFID and name	Recall
Rheinisch-Westfälisches Institut für Wirtschaftsforschung e.V.	60029848: Rheinisch-Westfälisches Institut für Wirtschaftsforschung Essen	0.03
Technische Hochschule Mittelhessen – THM	60000512: Technische Hochschule Mittelhessen	0.04
Fraunhofer MEVIS	60107767: Fraunhofer-Institut für Bildgestützte Medizin MEVIS	0.06
Hochschule Braunschweig/Wolfenbüttel, Ostfalia Hochschule für angewandte Wissenschaften	60028838: Ostfalia Hochschule für angewandte Wissenschaften- Hochschule Braunschweig/Wolfenbüttel	0.08
Katholische Hochschule Nordrhein-Westfalen - Catholic University of Applied Sciences	60006764: Katholische Fachhochschule Nordrhein-Westfalen	0.08

The direct comparison of the results for the indicator scores, calculated with the Scopus platform disambiguation system –AFID– on the one hand, and those calculated with the KB system on the other, in terms of distributions of percent deviation, are given in Table 4.12 and the deviation distributions are displayed in Figure 4.6. We find on average, for the absolute indicators, considerable shares of scores that are outside the range of accepted values. Relative indicators scores are less severely affected, but not within the accepted range often enough to be

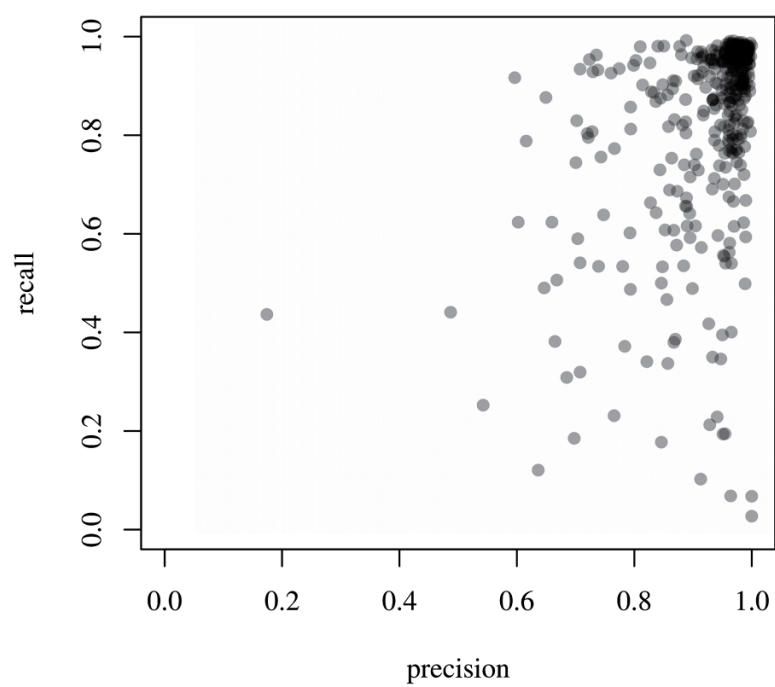


Figure 4.5: Precision and recall for Scopus AFID ($n = 376$).

considered reliable. It is worth pointing out that in particular the total number of citations (TCS) is rarely within the allowed range, which, however, did not seem to overly affect the other citation indicators.

Table 4.12: Deviation of indicator scores of Scopus AFID from KB system in Mode A ($n = 376$)

Indicator	Percent deviation within ± 5 %	Median absolute deviation [%]
P	40.7	9.7
P(collab)	40.7	9.3
PP(collab)	86.7	1.4
P(int collab)	40.8	8.9
PP(int collab)	71.5	2.7
TCS	12.8	15.0
MCS	59.0	4.9
TNCS	41.0	9.7
MNCS	66.5	3.1
P(top 10%)	40.2	8.6
PP(top 10%)	68.1	3.1

4.5 Discussion

We have investigated the accuracy of bibliometric indicator values for German publicly funded research organizations that can be obtained through a search strategy on vendor-normalized data (for WoS) and through the use of the database vendors' proprietary institution disambiguation systems (for both WoS and Scopus). These indicator values were compared with results from a nearly complete and independent institutional disambiguation for which detailed performance characteristics were provided.

During our study, we found that conceptual differences between the three institution disambiguation systems and a lack of documentation of both the WoS OE system and the Scopus AFID system were obstacles to making straightforward comparisons. In particular, the definition of the basic institutional entity—which is a crucial point for comparing disambiguation systems—varied among the systems. For example, in Scopus, university hospitals were kept separate from university entities. They had different AFIDs, which were not connected in any way. This inhibits evaluations for universities including their academic hospitals or medical faculties. For a comparison with the KB system, these entities, academic hospitals and the universities to which they belong, had to be

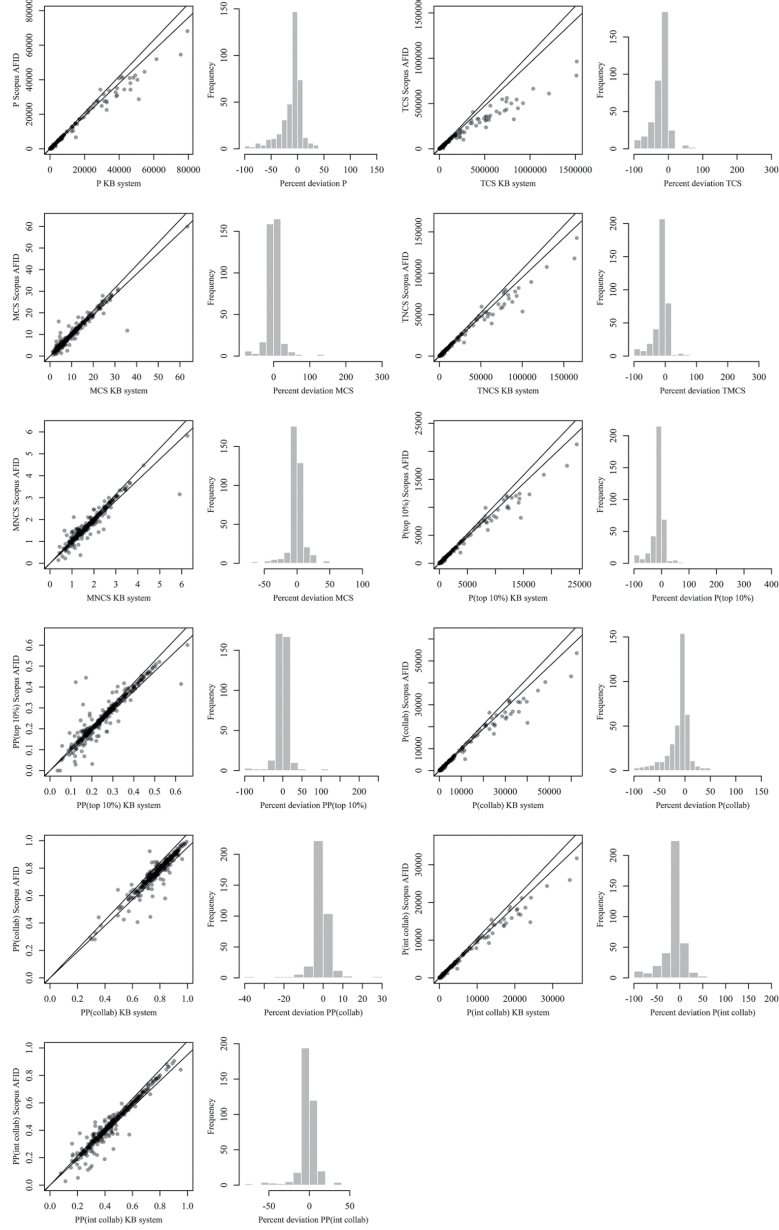


Figure 4.6: Distributions of indicator score deviations of Scopus AFID from KB system in Mode A. Diagonal lines indicate the $\pm 5\%$ error margin for indicator values.

aggregated manually. A further issue was faced regarding the handling of predecessor institutions. In order to obtain valid results, we evaluated the systems on their own terms, adjusting the KB system as necessary, to include predecessor institutions. In WoS OE, the level at which institutional entities are defined (e.g., MPG as one single institutional entity), largely rules out a comparison on the institutional level, as defined in the KB system, for some KB sectors. Furthermore, there is no clear documentation on the handling of structural changes over time, such as splits or mergers. For analyses at the institutional level, this is a major limitation.

In the WoS OE and Scopus AFID systems, precision of the obtainable publication sets was close to adequate levels at 0.95 and 0.96, respectively. However, neither system provided high recall rates (WoS: 0.93; Scopus: 0.86), which led to inaccurate indicator scores. Furthermore, we find substantial variation in precision and recall across institutions, indicating that within one system, these values are not systematically similar across the covered institutions but differ on a case-by-case basis. As for the tested name search method on normalized WoS data, precision and recall scores are poor, so this approach does not constitute a viable alternative.

Our results show that indicator values will typically not be within tolerable error margins at the organizational level, which we have set at ± 5 % of the reference value. This holds both for size-dependent and size-independent indicators. Hence, bibliometric indicator values at the institutional level have only limited accuracy.

Relying on vendor disambiguation systems may incur serious inaccuracies in indicator values at the institutional level. Therefore we conclude that for any use with policy implications, additional data cleaning for disambiguating affiliation data is recommended. We stress that any study such as the one presented in this paper shows only the current situation and that disambiguation systems may improve over time. The lack of adequate documentation of vendor institution disambiguation systems, including performance figures, is, however, another barrier impeding the adoption of these institution disambiguation systems in bibliometric studies.

Author contributions

Paul Donner: Conceptualization, Methodology, Investigation, Formal Analysis, Visualization, Writing—original draft, Writing—review & editing, Funding Ac-

quisition. Christine Rimmert: Conceptualization, Methodology, Investigation, Writing—original draft, Funding Acquisition. Nees Jan van Eck: Methodology, Investigation, Writing—review & editing, Supervision.

Competing interests

Paul Donner applies the KB institution disambiguation system for WoS and Scopus data in research and commercial projects. Nees Jan van Eck is affiliated with the Centre for Science and Technology Studies (CWTS) at Leiden University, which has developed its own affiliation disambiguation system for WoS data and which uses this system in research and commercial projects.

Funding information

This research was supported by Federal Ministry for Education and Research, Germany, project 01PQ13001.

Acknowledgements

An earlier version of this paper was shared with Clarivate Analytics and Elsevier. We would like to thank Clarivate Analytics and Elsevier for their feedback. This study was partially funded by German Federal Ministry of Education and Research (BMBF) project 01PQ17001.

Appendix A: Evaluation of the KB disambiguation system for German research institutions

A manual evaluation of the KB disambiguation system for German research institutions was conducted before the main study in order to assess its basic performance parameters and to inform any subsequent interpretations of results. The disambiguation quality is expressed in the information retrieval evaluation metrics of precision and recall. We begin by estimating the precision of the KB system for the data used in the study. In this task, precision is the proportion of correct document-address-combinations (assignments) to all assignments. From each result set of assignments produced by the KB disambiguation procedure for both WoS and Scopus data, a random sample of 1,000 assignments was checked manually for correctness of the mapped institution. The WoS sample contains 984 correct, 13 incorrect and three ambiguous assignments, leading to a precision of 0.9885 ± 0.0015 . The Scopus sample contains 981 correct and 16

incorrect assignments and three ambiguous cases—leading to precision of 0.9825 ± 0.0015 . We find precision to be similar for the two sets at approximately 98 %.

We now consider the estimation of recall, which is more involved. Recall, in this task, is the proportion of correctly assigned addresses to all relevant addresses of an institution in the data. Because the real proportion of addresses belonging to the institutions considered here is unknown, it is not possible to directly calculate recall for the specific context of this analysis. Therefore, we first compute the recall for the total sets of German addresses in WoS and Scopus for publication years from 1996 until 2013. From this basis, we will extrapolate to the subset used in the study in a second step.

Overall recall is calculated as the number of assigned German document-address combinations divided by the number of all German document-address combinations. The exact figures are stated in Table 4.13.

Table 4.13: Assigned vs. all document-address combinations (recall), German addresses 1996–2013

Document-address combinations of institutions (total set)	WoS	Scopus
All	3,178,143	3,530,460
Assigned to German research institutions in KB system	2,891,537 (90.88 %)	2,908,500 (82.38 %)
Overall recall	0.91	0.82

The difference between WoS (0.91) and Scopus (0.82) in recall probably arises because the WoS addresses were processed with higher priority than the Scopus ones. Furthermore, there is an issue with incomplete address records. There are instances where only the town and country, but no institution name, are provided in an address. We found more of these cases in Scopus than WoS, but this preliminary observation requires further examination.

Because our main study only concerns a subset of all institutions, namely higher education institutions and nonuniversity research institutes, we are interested in an estimate of recall that is more specific to this subset. One way to obtain such an estimate is to look into the document-address combinations that were not successfully disambiguated and thus will not be retrieved. More specifically, we are interested in the share of records of the selected types of institutions among the nondisambiguated records. For this purpose, random samples of 100 addresses without assignments were checked manually, using both the WoS and Scopus data, to determine whether they belong to the research institutions of

interest for this analysis. In the case of Scopus, 11 of 100 unassigned addresses belonged to institutions considered in this analysis; 86 addresses belonged to other institutions, with three ambiguous cases. The WoS sample contained 12 addresses of institutions of interest; 84 addresses belonged to other institutions and six cases were ambiguous. From these results, we obtain the following shares of the types of institutions of interest, in the set of not disambiguated records, including the ambiguous cases. WoS: 12 + 6 out of 100, a relative share of 0.18; Scopus: 11 + 3 out of 100, a relative share of 0.14. As the next step, we use these shares as factors for weighting the number of unassigned records to obtain estimates of the total numbers of records which cannot be recalled.

Recall for the selected set can then be estimated as the number of assigned document-address combinations in the selected set, divided by the number of all relevant document-address combinations in it. The unknown denominator –the total number of relevant records– is estimated on the basis of the sample described above: number of assigned addresses in the selected set + estimated number of addresses of interest in the set of unassigned addresses (derived from the sample).

We illustrate this calculation for the case of the WoS. The number of assignments for the types of institutions and publication years considered in the study is 2,211,795, which is a subset of the 2,891,537 total WoS assignments. This is the figure for the relevant recalled records – the numerator of recall. The number of relevant records –the denominator of recall– is not exactly known, but is estimated based on the checked sample. Of the 286,606 German address records not assigned to any institution, approximately 18 % belong to the types of institutions considered in the study. Therefore, we arrive at an estimated number of relevant records of 2,263,384. The ratio of assigned relevant addresses to the estimated total relevant addresses is the estimated recall value specific to the publication set of the study; in this case, the value is 0.977.

The recall value for the Scopus data, following the same method, is 0.962, see Table 4.14.

Table 4.14: Assigned vs. all document-address-combinations (recall), selected set

Document-address combinations of institutions in the selected set	WoS	Scopus
All (estimated)	$2,211,795 + (286,606 \times 0.18) = 2,263,384$	$2,217,031 + (621,960 \times 0.14) = 2,304,105$
Assigned	2,211,795	2,217,031
Estimated recall for the selected set	0.977	0.962

5 Data inaccuracy quantification and uncertainty propagation for bibliometric indicators

This chapter was first published under the same title in *Research Evaluation*, vol. 33, paper ID rvae047, at <https://doi.org/10.1093/reseval/rvae047>, on 09 November 2024.

Abstract

This study introduces an approach to estimate the uncertainty in bibliometric indicator values that is caused by data errors. This approach utilizes Bayesian regression models, estimated from empirical data samples, which are used to predict error-free data. Through direct Monte Carlo simulation – drawing many replicates of predicted data from the estimated regression models for the same input data – probability distributions for indicator values can be obtained, which provide the information on their uncertainty due to data errors. It is demonstrated how uncertainty in base quantities, such as the number of publications of certain document types of a unit of analysis and the number of citations of a publication, can be propagated along a measurement model into final indicator values. Synthetic examples are used to illustrate the method and real bibliometric research evaluation data is used to show its application in practice. Though in this contribution we just use two out of a larger number of known bibliometric error categories and therefore can account for only some part of the total uncertainty due to inaccuracies, the latter example reveals that average values of citation impact scores of publications of research groups need to be used very cautiously as they often have large margins of error resulting from data inaccuracies.

Keywords

data quality; citation analysis; uncertainty propagation; bibliometric data

5.1 Introduction

The formal analysis of uncertainty in results due to data errors is rarely considered in quantitative science studies and evaluative bibliometrics. We know only very little about how accurate and robust scientometric measurements and results are in light of pervasive data errors. Knowing the measurement quality can lead to an improved understanding of the fundamentals of the science

system and its dynamics: when we become aware of the measurement error of instruments and methods, we can better appreciate their affordances and limitations, and therefore the limitations of the results that were obtained with them. A crucial aspect of measurement quality in scientometrics is the quality of the original data. Sufficient data quality of third-party data is routinely taken for granted and rarely scrutinized independently, even though a considerable number of studies have brought to light severe and persistent levels of error in basic bibliometric data. Known data quality issues are usually not integrated into research, for example by sensitivity analysis or reporting error margins for result values. The bibliometrics and research evaluation community is aware of this deficiency to some degree. For instance, the influential Leiden Manifesto for research metrics (Hicks et al., 2015), in principle 8, establishes the following straightforward precept: ‘If uncertainty and error can be quantified, for instance using error bars, this information should accompany published indicator values. If this is not possible, indicator producers should at least avoid false precision.’ Nonetheless, it is exceedingly rare to come across bibliometric studies which include quantification of uncertainty due to data errors. In part, this is likely because of a lack of an accepted method to calculate and express these uncertainties.

In contrast, it is a widely accepted convention of the physical sciences that the result of a measurement consists of the best estimate of the value of a quantity and the uncertainty of that estimate (Taylor, 1997), for example, as in the relative uncertainties of measured fundamental physical constants in Tiesinga et al. (2021). Reported uncertainties are quantified statements on the quality of a measurement and allow users to judge its reliability. Known uncertainties of input quantities can be propagated along measurement models to obtain the uncertainties of output quantities by well understood procedures, called uncertainty propagation (JCGM, 2008; Taylor, 1997).

It is possible to apply this framework of uncertainty and its propagation to the social sciences, even though they are not dealing with measurement in the physical sense. As an example for translating this concept to scientometrics, consider the following. The citation counts of individual publications recorded in bibliometric databases are not always accurate. These errors can be modelled and described by a probability distribution. The uncertainties of the citation counts of a number of individual publications and a specified indicator formula determine the uncertainty of the value of that citation impact indicator, such as

the average number of citations of the publications of a research organization. The indicator formula or algorithm corresponds to the measurement model of physical science uncertainty propagation. If the distribution of the citation count error is known and can be formalized in a stochastic model, its influence on indicator results can be quantified by simulation studies using the Monte Carlo method. How is this useful? For instance, if the difference of the measured citation impact scores of two compared units of assessments is on the order of 10 per cent, it is important to know if the uncertainties of the two compared indicator result values are on the order of 2 per cent or 20 per cent. Only in the first case can an interpretation of the observed difference in the scores be justified as reliable. Uncertainty estimates based on data error distributions permit a judgment on whether the available data is accurate enough to achieve valid, robust, and useful results. This is the rationale for studying uncertainty propagation for bibliometric data and indicators.

The approach to uncertainty propagation for bibliometric indicator values proposed in this paper can be summarized abstractly by the following sequence of steps.

1. Define the types of error or data quality issues whose influence on bibliometric scores is to be quantified.
2. Collect an appropriate empirical sample. Determine the correct values of the variables in question in addition to the observed, possibly erroneous values.
3. Specify a Bayesian regression model structured to predict the error-free values of the variable from the observed values and any relevant available covariates. Estimate the model with the collected empirical data.
4. Use the estimated model for Bayesian posterior prediction of the likely error-free values based on the observed values of the variable in new data on the micro level of individual items a large number of times (Monte Carlo simulation).
5. Calculate bibliometric scores of interest from the predicted values for each Monte Carlo iteration.
6. Summarize and report the obtained indicator score distributions.

This approach is exemplified in the following sections for two sources of error, omitted citations and errors in document type assignments. As discussed below,

besides these there are many more error sources which are not included in this study, which means the expressions of uncertainty reported here only reflect a certain part of the total uncertainty arising from data inaccuracies. Such total uncertainties integrating all known such sources of data errors would realistically be of greater magnitudes, thus this study's uncertainty quantifications should be understood as partial and likely underestimating the full uncertainties.

The rest of the paper is structured as follows. In the next section the literature on uncertainty in scientometrics and on errors in bibliometric data is reviewed. Following that, we present empirical sample data for omitted citations. This data is then used in section 5.4 to estimate Bayesian regression models of the error distribution of citation counts. This section also introduces the proposed approach in a systematic manner. The stochastic error models are used to stage example simulations that illustrate how practitioners and researchers can use this approach to estimate the influence of data quality issues on study results. Conclusions are discussed in the final section.

5.2 Related work

5.2.1 Uncertainty in scientometrics not related to data errors

The present object of investigation, the quantification of the uncertainty of indicator values which is due to data errors, should be clearly distinguished from certain other concepts that are concerned with uncertainty and which also play a role in scientometrics.

In cases when the investigated objects form a well-defined population and it is impractical or not feasible to analyze all objects, one can resort to statistical sampling. When a random sample is studied, the values of quantities calculated on this subset of the population do not precisely match the values which could be calculated for the same quantity for the whole population, which is the actual quantity of interest. It is therefore important to quantify how similar the sample and population values are expected to be, in other words, the uncertainty of the sample value in representing the population value, i.e. the statistical sampling error. This uncertainty depends on the specific sampling method and sample size. Sampling uncertainty can be quantified, for example, by frequentist statistical confidence intervals for estimates. Some examples of scientometric applications of this approach are Schubert & Glänzel (1983) and Schubert et

al. (1988). Sampling uncertainty conveys important information on how well a random subset can be expected to represent a population based on insights from probability theory. It is not concerned in any way with errors in the values of data.

Waltman et al. (2012) introduced stability intervals for values of bibliometric indicators. These intervals are intended to quantify the relative susceptibility of an indicator value to minor changes in the underlying publication data, in other words, they reflect how stable or volatile the actually observed value is with respect to the observed data. If a bibliometric indicator value is highly dependent on relatively few input values, its stability is considered low and its stability interval is relatively large. Stability intervals are operationalized by means of bootstrapping, that is, drawing repeated samples with replacement from the available data, where the resampled number of observations is equal to the original number of observations, and computing the statistic of interest for each sample. The width of the bootstrap intervals of statistics can be computed in various ways from the distribution of resampled values of the statistic of interest (replicates), most simply by using the 0.025- and 0.975-quantiles of the bootstrap distribution to specify the 95 per cent bootstrap confidence interval, which is the method preferred by Waltman et al. (2012).

This approach has some conceptual difficulties. Because the bootstrap method is intended for statistical inference from random samples of larger populations (Efron & Tibshirani, 1993, p. 17), it fundamentally is based on the assumption that the set of observations could have been different. Bootstrapping, being a non-parametric method, frees one of the necessity to assume a specific distribution of the values in a sample and can be used to do inference for problems for which no analytical solution is available. Yet it does not do away with the prerequisite of sampling. This assumption of random sampling from a larger population is generally not met in bibliometric applications, which is one of the reasons why null hypothesis significance testing in research evaluation was criticized by Schneider (2013). When the stability interval is calculated for a bibliometric indicator summarizing the citation distribution of a research institution or journal, the entire population of publications can usually be observed or otherwise the observed subset is not a random sample and it is not clear how the set of observations could have been different – certainly not by drawing another sample from the same population. Moreover, resampling with replacement on such a data set suggests that publications could have not been produced or

could have been produced more than once while they were in fact produced exactly once. Resampling with replacement introduces new data which was not there and leaves out other data which was actually present.

The stability of bibliometric indicator values which are averages is related to the number of observations – more observations lead to more stable values – and the skewness of the distribution – more skew leads to more volatile values. Notwithstanding the above conceptual reservations, stability intervals have been shown to be of practical utility as they do in fact show quite clearly those values which are greatly affected by few unusual observations as the examples in Waltman et al. (2012) demonstrate. This concludes our brief digression on sources of uncertainty unrelated to data quality.

5.2.2 Accuracy and completeness of citation links in bibliographic databases

Bibliometric data from commercial data sources such as Web of Science (Birkle et al., 2020) or Scopus (Baas et al., 2020) and non-commercial sources like OpenAlex (Priem et al., 2022) is widely used in both academic science studies and in applied research assessment, monitoring, and funding distribution. An example is the Science & Engineering Indicators report series of the U.S. National Science Foundation’s National Science Board, which has used Science Citation Index data for many years and switched to Scopus data in 2016. Similar monitoring reports are produced with bibliometric data in other countries (Moed, 2005, p. 271). Performance-based funding systems in many countries include a bibliometric component (Hicks, 2012; Zacharewicz et al., 2019). Various global university rankings base their scores on bibliometric data. For instance, the Times Higher Education and the QS rankings currently use Scopus data, while ShanghaiRanking and U.S. News & World Report use WoS data.

Bibliometric data sources, including commercially produced ones, are not error-free (e.g. Prancutė, 2021, pp. 20–24). There are a number of types of errors or inaccuracies in bibliometric data that can exert adverse influence on the results of bibliometric studies. In this paper, only two types are studied: incomplete citation counts and incorrect document type assignments. Other types of errors include, but are not limited to

- incompletely or incorrectly identified authors and institutions (Dickersin

- et al., 2002; Donner et al., 2020; Schulz, 2016),
- incorrect or missing publication identifiers (Franceschini et al., 2015a),
- duplicate records (Liu et al., 2021; Valderrama-Zurián et al., 2015) and missing records (Krauskopf, 2019),
- insufficiently standardized publication venue data (Leydesdorff et al., 2016, pp. 2134–2136),
- ignored citation data in supplemental information, a type of missing citations (García-Pérez, 2015),
- incomplete or erroneous funding data (Álvarez-Bornstein et al., 2017; Liu, 2020)

and other missing data elements such as classification assignments of publication records.

Citation counts of scientific research contributions are a fundamental type of data in scientometrics. Accuracy and completeness of citation links are therefore crucial data quality issues (Moed, 2005, ch. 13; Olensky et al., 2016, pp. 2550–2551). It has been demonstrated that citation counts in bibliometric data sources are subject to non-negligible error rates in a number of studies.

Moed & Vriens (1989) studied causes of unmatched references to a sample of about 4500 target records from five journals in an online version of the Science Citation Index database. Note that in this study, the reference matching procedure of the Science Citation Index was not studied at all. Rather, it was investigated why for some cited references no exact match could be found even though the cited journal and year of the cited reference were covered in the data. Selecting citing references by cited publication year and cited journal name, including variations, they focused on the data fields cited first author name, volume number, and starting page. They found that about 9 per cent of references contained at least one discrepancy in these fields as compared to the identified corresponding target record. Some particular journal conventions, such as ‘combined volumes’ and special pagination for letter sections, could be identified as associated with certain error types, as were some types of author names, particularly those consisting of more than a single word. In auxiliary analyses the authors also established that discrepancies between publication year of the target item and cited reference are common. The results demonstrate convincingly that any strict reference matching using only exact agreement of field values would overlook a considerable number of citation links and that there-

fore reference matching algorithms in general need to be lenient with respect to minor variations in field values between cited and citing record in order to attain high completeness of citation counts. Moreover, there are clear systematics at play both for entire target journals and for individual publications, which precludes any assumption that these errors can be neglected because they are distributed randomly and affect all cited documents equally.

Another early empirical study is that of Abt (1992) who looked up around 1000 cited references from one astrophysical journal issue to eight cited journals using the original volumes or published indexes of the cited literature. Some 12 per cent of cited references were found to be inaccurate in the original reference lists. Abt found that most of these were corrected during processing for Science Citation Index: ‘the 12.2% author errors result in only 3.6% misplacements or omissions in SCI’ (p. 236).

Buchanan (2006) investigated the accuracy of cited references in two citation databases, Science Citation Index Expanded (SCIE) and SciFinder Scholar, using as source data some 6000 cited references from around 200 articles in three chemistry journals. Checking both the original publications’ reference lists and the indexed versions allowed him to attribute the cause of found errors to either the authors or to processing by the citation databases. In both databases, around 10 per cent of cited references were not matched to their indexed target. The major reason were author errors but database mapping errors were a considerable factor as well. Among SCIE database mapping errors, the predominant type was omission of cited references. This study confirmed systematic errors for specific journals and individual articles.

Franceschini et al. (2013) propose a method to study the error rate of citation links at a large scale with an automated approach by comparing sets of citing and cited papers that are covered by two or more databases, similar to the earlier short study of Dyachenko & Pislyakov (2010). As it is automated, the advantages of the method are that it can be applied directly to any set of publications of interest and that it provides individual paper-level error estimates. Drawbacks are that, first, subscriptions to multiple databases are required and second, that, because the shared citing publications are not independent, it is likely that for cited references with severe inaccuracies, more than a single database would fail to link it to the intended target. If all studied databases cannot resolve a cited reference, the method cannot detect that specific error. Using as a test case articles from three journals from publication year 2008 and

the two databases Web of Science (WoS) and Scopus, the study finds error rates of 3.2 per cent and 5.6 per cent, respectively, for this data set. Furthermore they propose a statistical model for estimating the true number of citations from the empirical error rate and the number of observed citations.²⁶

In the first study of its kind, Olensky et al. (2016) studied the performance of two proprietary reference matching algorithms on WoS data compared with the results of the algorithm of WoS itself. Because of the prevalence of inaccuracies in cited reference data and the conservativeness of the WoS algorithm, the two proprietary algorithms exhibited substantially higher Recall at the cost of slightly lower Precision in a sample of about 4000 citations to 300 target papers. While the WoS algorithm missed about 6 per cent of reference matches, the two other algorithms missed about 1 per cent each. These outcomes show that the WoS algorithm is apt to make no or very few false positive errors but overlooks inaccurate references leading to false negatives, which can be overcome with alternative matching strategies. The study also found a number of citation links present in the data which were not based on cited references in the original publications.

Van Eck & Waltman (2017) studied the representation of a large data set of cited references of papers published in journals by the publishing house Elsevier in the citation databases WoS and Scopus. One important finding is that in WoS the number of cited references is lower than that of the original publication for 19 per cent of matched papers while the corresponding figure for Scopus is only 1 per cent. Both databases had a similar share of papers without indexed references (about 1 per cent) which do have cited references in the original publications. Inspection of a sample showed that both databases are prone to miss cited references present in the original data.

Further studies concerned with cited reference data errors include Tunger et al. (2010), Schmidt (2018), Dyachenko & Pislyakov (2010), Olensky (2015), and Tüür-Fröhlich (2016).

While in the past decades citation matching had to rely on bibliographic meta-data, such as author names, source and publication titles, publication year and

²⁶As the equations of Franceschini et al. (2013) for the expected value (Eq. 12) and variance (Eq. 13) of the number of omitted citations o_i are multiplicative in c_i , the number of observed citations, their model has the limitation that any uncited publication has an expected value of zero omitted citations with zero variance. The same holds for their equations for these variables for sets of papers if the total sum of citations is zero.

volume number, now most publications are published with unique unambiguous identifiers, Digital Object Identifiers (DOIs). DOIs have also been assigned retrospectively to a large share of published scientific literature and they are widely accepted as part of reference lists. As they in principle permit an error-free matching of cited reference to a record of an indexed document, one might expect that the error in citation matching is much reduced for the more recently published literature for which citation database vendors can utilize DOI matching. However, the data of identifiers is just as much subject to data quality problems as any other data (Franceschini et al., 2015a, 2016; Krauskopf & Salgado, 2023; Zhu et al., 2019). For instance, C.-K. Huang et al. (2020) compared three bibliometric data sources and found some DOIs only present in one of them. Checking a sample of them manually, they report:

We have also checked each DOI against the DOI string actually recorded as per original document (where applicable) or via doi.org. These percentages (of correct DOIs) are 93.1%, 98.2% and 96.7 for WoS, Scopus, and MSA respectively (with all 15 institutions combined). While these numbers are relatively high, a significant number of errors suggests that DOIs are not being systematically checked against authoritative sources such as Crossref which we find surprising.

DOI errors have also been found to be prevalent in the cited references data (Cioffi et al., 2022; Xu et al., 2019), confirming that there data quality issues on both the citing and the cited sides of the reference matching task. Clearly, the introduction and widespread use of unique identifiers in cited reference data has not solved the citation matching problem.

From this cross-section overview of the literature the following conclusions can be drawn:

1. References in the original publications are often incorrect (Dubin, 2004; Jergas & Baethge, 2015; Moed & Vriens, 1989; Sauvayre, 2022; Wager & Middleton, 2008).
2. Errors in citation matching in citation indexes can have several causes. Besides the above item 1, a reference in the citation index can be an inaccurate copy of an original correct reference. The reference matching

procedure may be unable to resolve some surface form of a reference to another, that is, link a source citation to its intended target.

3. The proprietary reference matching procedure of WoS is based on close agreement of cited reference data (normalized and ‘re-written’ during processing (Garfield, 1983)) to the target papers’ bibliographical data, which is also normalized during source indexing. This corrects a considerable share of errors in cited references data made by citing authors (Abt, 1992; Buchanan, 2006). Consequently, the reference matching procedure of WoS has near-optimal Precision but incomplete Recall – it is known to miss some inaccurate reference links (Olensky, 2015).
4. Advanced reference matching algorithms that are more error-tolerant can identify more correct citation links in the case of WoS data as compared to WoS’s own algorithm.
5. The share of unmatched references that should ideally be matched is roughly around 5 per cent.
6. The indexing process of references by citation databases is susceptible to the introduction of errors such as overlooking references or wrongly correcting erroneous references. The scale of these inaccuracies is more difficult to determine but probably lower than the unmatched reference rate.

Despite these known flaws of reference matching algorithms and inaccurate references in the original publications, usually no attempts are made to incorporate this uncertainty about citation counts into the reporting of indicator results. In the second part of this paper we sketch a method of how this could be done.

What is also missing is an estimate of the rate of missed citations by a principled method for a random sample – in the sense of being representative for a complete citation index database. For the first part of this study, a simple random sample of WoS source papers was drawn and it was attempted to find all reference strings of WoS indexed documents that reference them. Considerable effort was expended on identifying inexact matches. The objective is to give a statistical estimate of the proportion of missed citations and to describe the relationship of the number of found citations to the number of missed citations, i.e. the conditional error distribution. While we are concerned with the specific case of WoS data in what follows, the problem of suboptimal reference matching is just as acute for other citation index databases.

5.3 Missing links: Empirical study of citation error distribution

5.3.1 Methods and data

In this section we report on an empirical study²⁷ of the prevalence of incorrectly matched references in WoS. Incorrectly matched references result in incorrect citation counts. The analysed data originate from licensed raw data in tagged data format of the WoS journal and proceedings citation indexes²⁸ as of spring 2015. A simple random sample of target items published between 1980 and 2015 was drawn from all journal articles, letters and reviews, as defined in the data. Hence, the specific results of this study apply to WoS, not other citation databases. They also do not apply to specific subsets of publications whose error structure is known or presumed to be different, such as publications with references in footnotes, which are more difficult to extract. For such cases, as for other databases, new empirical samples should be collected and analyzed.

Cited references data of all publications from 1980 to spring 2015 were indexed for search. The reference strings, consisting of author name field, split into last name and initials at the comma, the source title, publication year and first page fields were indexed with Oracle Text in an Oracle Database instance. The volume field was not considered because, for the target journal items, volume and publication year are nearly redundant information but publication year is more accurate and more complete in WoS reference data than volume (Moed, 2005, Table 13.1).

A procedure was programmed to search the index for references likely referring to the sampled target items. Because the search field data of the target items has to correspond to the way the reference data is prepared and stored in WoS, the target author name and source title were pre-processed. This entailed the abbreviation of first name to initials, the removal of non-letter characters in the name, and, for the source title, using the WoS abbreviation. Where more than one abbreviated title version for the same journal existed in the data and when an additional author group name was available, all possible combinations of

²⁷Initial results from this study were first reported in Donner (2016a). The dataset is available at <https://doi.org/10.5281/zenodo.13969973>.

²⁸Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Conference Proceedings Citation Index - Science, Conference Proceedings Citation Index - Social Sciences & Humanities. All licensed for records of publication years 1980 to present.

those fields' values were used as search input. The procedure performs a fuzzy search on the index and returns a list of unique candidate reference strings that are sufficiently similar to the target input. The search is deliberately lenient so that all possible matches are returned in order to prevent false negatives as much as possible, which is a requirement for this study. This procedure searched for simultaneous fuzzy hits on first author lastname, first author initial (target item co-authors were not considered), source title, publication year (formatted as string), and firstpage. This strategy found 19132 citation record candidates for 372 sample target items. Each record could stand for one or more citations as only unique records were searched and checked.

The candidates were checked for whether they constitute a match to the target or not by a student assistant. Ambiguous candidates were afterwards assessed by the author. Care was taken to avoid false positive matches by querying the database for any exact matches of the candidate reference strings other than the target item. This entire procedure was completely independent of the reference matching done by WoS. The found positive matches are used as citation links and the derived independent citation count for each target item is calculated by retrieving all references using those unique candidate strings.

The WoS-provided citation links were obtained for comparison. They are constituted by the WoS matchkey, designated as the T9/R9 fields in the tagged data files. All citations until 2015 were counted. Table 5.1 below shows one specific example, indicating which variants were found by the above method for this study and which were available in the original WoS data. Column "Identity Hits" indicates whether there is a source record in the database that matches exactly with the reference in all fields. The row "target source item" gives the specific form in which a sample publication should appear in the references data according to WoS formatting conventions. Row "citation 1" shows that this exact form (combination of family name, initials, year, source, first page) did in fact occur in the data 109 times and was classified as a citation to the target by both the WoS algorithm and our independent inspection. Of the other candidates, the WoS algorithm also counted "citation 4", a variant in which the publication year is incorrect and a variant of the source name abbreviation is used. This variant occurred twice. In total WoS found 111 citations for the target item. Besides "citation 4", we were also able to identify "citation 2" (year error), "citation 5" (first page error) and "citation 7" (missing first page, variant source abbreviation) as valid citations, which added up to four additional

citations. We judged “citation 3” as not valid (different year and first page) and “citation 6” as not valid (different author name initial, first page and volume; and matches another source record in the database exactly).

Table 5.1: Example of reference candidate records

Record	Family Name	Initials	Year	Source	First Page	Occurrence	Identity Hit	Match WoS	Match this study
target source item	SMITH	DS	1981	BRIT MED J	517				
citation 1	SMITH	DS	1981	BRIT MED J	517	109	yes	yes	yes
citation 2	SMITH	DS	1986	BRIT MED J	517	1	no	no	yes
citation 3	SMITH	DS	1969	BMJ-BRIT MED J	90	1	no	no	no
citation 4	SMITH	DS	1986	BR MED J	517	2	no	yes	yes
citation 5	SMITH	DS	1981	BRIT MED J	5	4	no	no	yes
citation 6	SMITH	R	1986	BMJ-BRIT MED J	1043	1	yes	no	no
citation 7	SMITH	DS	1981	BR MED J		1	no	no	yes

5.3.2 Results

A total of 372 sampled WoS records were assessed. According to WoS, these were cited a total of 6120 times. Our search found 255 additional citations, which means that 4 per cent of the citations were missed by the WoS matching algorithm, which is in agreement with the results of other studies (Buchanan, 2006; García-Pérez, 2010; Olensky et al., 2016). The distribution of missed citations is presented in Table 5.2. It is clearly apparent that the distribution of omitted citations is neither symmetric nor normally distributed. The inclusion of these citations would increase the WoS average citations per paper in the sample from 16.4 citations by 0.7 to 17.1. Of the 372 sample target items, 29.4 per cent have one or more detected missing citation. An association between the apparent citation count and the citation count error can be observed, as the number of WoS citations and the number of missed citations per item are moderately correlated with $r = 0.31$, 95 per cent confidence interval: $[0.22, 0.40]$, see also Figure 5.1. This means that from the perspective of cited items, missing citations due to unsuccessful matching are not random but structured in the sense that those with higher citation impact (more matched citations) also tend to have more missing citations.

Our study design did not allow us to identify any false positive citations, also called “phantom citations”, i.e. citation links present in the database where no corresponding cited reference was present in the original publication. This was not possible because we only used the data present in WoS. As mentioned above, WoS transforms the original data and only the transformed data can be accessed. Phantom citation errors have been found by other studies (Dyachenko & Pislyakov, 2010; García-Pérez, 2010, 2011; Olensky, 2015, pp. 114–115; Van Eck & Waltman, 2017) and are more rare than missing citations. They do constitute one cause of missing citations because they mis-link an original reference to an incorrect cited source record. Because we only used WoS-internal data we also could not assess the phenomenon of references that are not transferred from the original publication to the database record (Buchanan, 2006). This is another, different kind of missing citations than the one addressed in this paper. In our study we are restricted to indexed but unmatched missing citations. Finally, as we used the cited reference source data and compared it with the publication record source, we did not cover references to preprint or working paper versions of a publication, only to the publisher version.

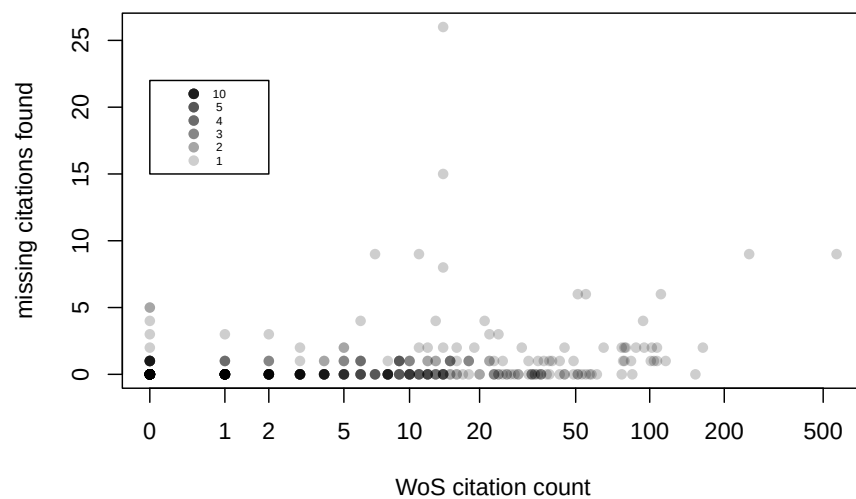


Figure 5.1: Scatterplot of WoS citation count and number of missing citations found

Table 5.2: Distribution of number of missed citations from 372 records in WoS

missing citations	frequency
0	263
1	67
2	20
3	5
4	5
5	2
6	3
8	1
9	4
15	1
26	1

5.4 Propagation of uncertainty from data to bibliometric indicators – a Bayesian regression approach

5.4.1 Incorporating information about uncertainty into statistical models

In the preceding section we have empirically determined the magnitude of one particular source of error of bibliometric data, missing citations due to incomplete reference matching. Consider that we might also collect other samples for other error sources. Having obtained such empirical data on error distributions, it becomes possible to apply this information to other, out-of-sample data to estimate the effect of data inaccuracies on indicator values of interest for this new data. In this section, we propose a methodology for using empirical data on bibliometric error distributions for obtaining uncertainty intervals for bibliometric indicator values on new data from the same data source with the same properties of data inaccuracies. For this purpose we construct statistical models of the error processes which allow to abstract and generalize from the specific empirical sample data so that one can make stochastic predictions of the likely errors in any new data. Such regression models are estimated once and predictive simulations from the models for other data can be used to specify the uncertainty, not only at item micro-level and the average level across papers, but also at the level of specific indicator values at any aggregation, to quantify how precise indicator values are, given known uncertainty structures in the underlying data. This can be done by using the regression models to make statistical predictions for error-corrected values for new input data, such as the likely error-free citation count predicted from observed error-affected citation count values. Making such simulations a large enough number of times for each input observation gives a probabilistic approximation of the distribution of likely error-free outcome values.

This approach is called the Monte Carlo method. In a Monte Carlo simulation, a process is simulated by repeated random sampling of input data to obtain probability distributions of some outcomes of interest. The random variation is a representation of the uncertainty in the modelled relationships. The shape of the probability distributions of the outcome variables reflect the uncertainty in the simulated results. The simulated outcome values' probability distributions after sampling a sufficiently large number of times can then be statistically

described and analysed.

It is possible to construct two types of statistical models from empirical data that can be used to understand the effect of data quality on bibliometric indicators. They differ by the direction of prediction: one can either predict erroneous data from error-free data or the reverse. In the models of the first kind, we start with a given data set without any errors. This could either be real data that was thoroughly checked for quality and corrected by hand, or simulated data. We then use a statistical model of the error process, obtained from other empirical sample data, to predict the degree to which empirically observed levels of data quality issues affect bibliometric indicators calculated from the data. This kind of model is useful for testing hypothetical scenarios, such as ascertaining the error rate in raw data that can still be tolerated if one is interested in comparisons of some indicator scores with a specified accuracy. For example one might create a data set of simulated bibliometric ground truth data and use a model of the first kind to test what range of values would be expected if this data is subjected to the process of indexation in bibliometric databases that results in the observed inaccuracies.

An exposition of the use of the models of the first kind is relegated to the appendix in order to keep the main body of the paper focused on what we consider the more immediately and broadly useful practical applications.

In the models of the second kind, we work with given data from a bibliometric data source for which we know the general relevant error rates from earlier empirical error studies but not the specific inaccuracies in the data at hand. We then use prediction from estimated statistical models to generate a large number of simulated hypothetical error-free versions of the data, which are consistent with the given data and the knowledge about error distributions, and calculate hypothetical versions of the relevant bibliometric indicators from the simulated data. The distributions of the simulated indicator values contain the information about the uncertainty of the indicator values which can be summarized, such as with point estimates accompanied by 95 per cent probability credible intervals as error margins of error-free indicator values. That means that these models work in the inverse direction from the models of the first kind. While the models of the first kind start with error-free data and simulate realistic errors, the models of the second kind start with error-affected data and simulate realistic error-free data. Both kinds of models can be estimated from the same empirical data sets, only the dependent and independent variables are switched.

5.4.2 A simulation exercise of bibliometric uncertainty propagation – Simulating error-free data

The typical situation is that researchers have to work with data which is known to be inaccurate but for which it is not feasible to make full corrections at scale. With models of the second kind as introduced above, we can utilize limited samples of empirical data on error distributions to stochastically predict probable correct values for newly observed data. By repeated predictions of hypothetical correct values with random variation at the level of attributes of individual publications simulated distributions of indicator values can be calculated from the predicted basic data and summarized and expressed as uncertainty intervals for bibliometric indicator values, be it at the level of individual publications or arbitrary aggregates.

Above, we presented data on the uncertainty in citation count figures for one specific data source, WoS. There are other sources of uncertainty and in the following we will also consider one further source of uncertainty, namely errors in the document types of items assigned by bibliometric databases.

To study the effect of data inaccuracies, statistical models of the error distribution are required. Here we propose to use Bayesian regression, primarily because this allows the uncertainty in the fitted model parameters (sampling error) to affect the predicted values for new data. This is possible because, unlike in frequentist regression, the parameters in Bayesian regression are not modelled as scalar values but as probability distributions. We use the empirical data from the previous section²⁹. This and all other models are independently estimated with the R package `brms`, v. 2.20, (Bürkner, 2017).

We estimated two independent Bayesian regression models. For the first one we estimated the number of missing citations of a publication from its observed citation count with a negative binomial model according to $o_i \sim \text{NegB}(\ln(c_i + 1), \theta)$ and predicted error-free citation counts calculated as $c_i^{*sim} = c_i^{new} + \hat{o}_i$. This notation follows Franceschini et al. (2013): in model estimation, o_i is the number of omitted citations and c_i the number of error-affected citations

²⁹The citation count error models are estimated on empirical data of only a moderate sample size. The sample does not include sufficient data at large observed WoS citation counts, a consequence of the basic sampling strategy and the skewness of citation count data. We therefore restrict our simulations to moderate citation counts. The collected data can be extended with new sample data and new models estimated as the need for such modelling arises.

of a publication, collected from a sample of empirical data; θ is the negative binomial overdispersion parameter, estimated from the data. In prediction from the estimated model, c_i^{*sim} is the simulated, i.e. predicted, error-free number of citations which is calculated from observed new citation count data c_i^{new} and predicted missing citations $\hat{c}_i$. A negative binomial model is used as it is suitable for non-negative and skewed count data. A log-transformation is applied to the independent variable as the otherwise excessive range would make the model unstable for high citation counts. To estimate the model, the data from section *Missing links: Empirical study of citation error distribution* are used. Once this model is estimated, it can be used to stochastically predict the number of missing citations for any desired new input citation counts in a way that fully incorporates the available empirical knowledge, in this case the overall incidence of omitted citations and the relationship between omitted citations and total citation count.

In Bayesian modelling empirical knowledge beyond the data used to estimate the model can be included by specifying prior probability distributions. These reflect prior knowledge about model parameters and get updated via Bayes' rule with the data to obtain the posterior distribution. In the case of missing citations, some prior knowledge is available in the form of estimates of the rate of missing citations due to inexact reference matching reported in earlier studies. The results of Abt (1992), Buchanan (2006), Franceschini et al. (2013), and Olensky et al. (2016) suggest that the rate of missed citation in WoS can be expected to fall somewhere between 0 and 12 per cent. Assuming a plausible central value of 6 per cent and given that the observed average citation count in our sample is 16.4, this would indicate an average value of 1.0 missed citations with values above 0.0 and up to 2.0 still being plausible. From this we set a weakly informative prior distribution for the intercept term of Normal(0, 0.8) on the log scale. This sets the prior distribution on the original data scale centered on 1.0 citations with 95 per cent of the probability mass between 0.2 and 4.95 citations. As we have no particular previous knowledge on the influence of the predictor, we set a vague Normal(0, 1) prior for its coefficient. Prior predictive checks showed that these settings generate distributions broadly compatible with the data.

For the second model, the error-free document type is predicted with a multinomial regression model in which the error-affected observed document type is the single predictor variable according to $d_i^* \sim categorical(d_i)$. The empirical data

to estimate this model is from Donner (2017). In that study, the true document type was collected by inspection of the original document and assigned according to the published definitions of document types by the two databases which were analyzed. Both document type variables have four classes: article, review, letter, and other. Before we proceed with simulation exercises which investigate the influence of the two error sources we outline how inaccurate document type data can adversely affect bibliometric indicators.

First, bibliometric studies normally consider a subset of all publications of a unit of assessment. This subset is usually defined based on publication years and document types. Often the document types article, review, and letter are chosen and all others are not included in an analysis because not all types of documents are of interest in bibliometric research or evaluation. Hence, if the document type assignments are not correct, it is possible that not all relevant publications will be selected for analysis and that some irrelevant publications will be included, which can lead to incorrect publication counts. In this scenario, the citation counts of wrongly left out and wrongly included publications also distort citation impact indicator values.

The second and third kinds of error which can be caused by inaccurate document type data arise when constructing normalized citation impact indicators using reference sets to obtain reference values. The second error arises because the reference values for normalized citation indicators are also based on the selection of publications of specific types. For example, an observed citation count for a publication of type article is compared to the citation count of all publications of document type article in the same field and publication year. If, due to document type errors, the wrong publications are inadvertently used to calculate these citation score reference values, the calculation of normalized citation scores for the publications of interest in a study will also be affected.

The third kind of error is that when the document type assignment of an item is incorrect, the normalized citation scores of the target publications might get calculated on the basis of the wrong normalization value, namely that of the reference set for the incorrect document type.

Simulation exercise 1 We begin with a deliberately simple example of the proposed method. Subsequent simulations will be incrementally more complex in order to develop understanding of the approach step by step. In this basic ex-

ample we look at the distributions of predicted error-free values of citation count and document type of three publications – only these attributes at the micro-level of publications are examined and no indicators are calculated. We are given as observed, possibly error-affected data, publications P1 of type article with 5 citations, P2 of type review with 10 citations, and P3 of type letter with 0 citations. Using these observed values and the regression models described above, we predict the likely error-free values of their characteristics with 2000 draws from the posterior predictive distributions of the models for each publication and attribute. Table 5.3 shows the predicted error-free document types of the three publications. According to these figures, it is unlikely but possible that P1 is a different type than article. In contrast, for P2 and P3 it is quite likely that the observed document types were inaccurate as substantial probabilities for types different than the observed ones are predicted. The figures for predicted error-free citations counts are shown in Table 5.4. In each of the three cases, the most likely citation count is the observed one but the numbers in the higher counts indicate that higher error-free citation counts are also consistent with the modelled empirical data and given observations. For example, for P2, the probability that the observed value of 10 citations is accurate is less than 70 per cent.

Table 5.3: Simulation exercise 1. Predicted error-free document types of three publications

publication	Article	Letter	Other	Review
P1	1942	4	10	44
P2	252	1	65	1682
P3	738	1082	178	2

Table 5.4: Simulation exercise 1. Predicted error-free citation counts of three publications

publication	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	25	33
P1	0	0	0	0	0	1477	280	116	57	36	14	9	7	2	1	0	0	0	0	0	0	1	0	0
P2	0	0	0	0	0	0	0	0	0	0	1362	320	146	65	38	21	14	16	9	3	3	1	1	1
P3	1638	243	80	21	12	2	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Simulation exercise 2 We now present a series of three simulations in which we first look at the effects of each of the two sources of data inaccuracies in isolation and then at their combined effect. To begin, in this exercise we look at

citation count error only and assume that document types are free of error. To make the exercises both instructive and interesting, we model a basic evaluation scenario. Two research units with modest numbers of publications are evaluated with respect to their publication output and citation impact. The fundamental question in this scenario is whether, given the known data quality issues, the observed differences in scores are reliable reflections of actual performance differences or whether observed differences might be attributable to data inaccuracies alone.

For these next three exercises we first create a common data set of error-affected values for two research units, A and B, with 40 and 50 publications, respectively, of various document types. For calculating bibliometric indicators, only the subset of articles and reviews of those publications will be selected and counted (indicator publication count, P) while letters and items of other document types are discarded, as is common in bibliometric studies. Let the units only have published in one homogeneous research discipline and in the same publication year, so no further normalization for these factors needs to be applied. An additional reference set of publications consists of 200 publications of all document types. Document types of all three publication sets are sampled from the global document type distribution which, based on WoS data, is as follows: articles: 68 per cent, letters: 3 per cent, reviews: 4 per cent, other: 25 per cent. Citation counts of papers, the input quantity in this exercise, are generated from discretized lognormal distributions such that units A and B and the other publications each have their own typical citation impact levels, realized by setting the mean parameters of the distributions the citation count values are drawn from to different values (A: 0.8, B: 1.2, ref.: 1). Depending on the document type these parameters are multiplied with scaling factors as follows. Articles: 1; reviews: 1.5; letters: 0.2; other: 0.1.

For the two units A and B, the three basic bibliometric indicators, number of article and review publications (P), number of their citations (C), and mean normalized citation score (MNCS), are calculated. For the latter indicator, first, the item-level normalized citation scores (NCS) of the publications are calculated as the observed citation count divided by the expected citation count. The expected citation count is the average of the citation counts of all publications of the same document type. The MNCS of a unit is then computed as the arithmetic average of the NCS of its publications (Waltman et al., 2011). The choice of indicators is purely illustrative and the present approach can be used for any

other indicators.

For each of the observed citation count values of the whole publication set 2000 draws from the posterior predictive distribution of error-free citation count were drawn from the estimated citation count error Bayesian regression model. For each of these 2000 draws, the bibliometric indicator values for the two units were calculated. The distributions of these indicator values across the simulations are reported in summarized form in Table 5.5. In this and the following similar tables the original fixed error-affected indicator scores are displayed in the left part and the values for the distributions of simulated error-free indicators in the right part of the tables. As document types were assumed to be error-free, the publication count indicator P is not subject to any variation. The units' total citation counts (C) and mean normalized citation score (MNCS) show clear differences in their median values as compared to the input values. Probable error-free citation counts are considerably greater than observed values would suggest. MNCS values for simulated error-free data are close to the observed values in the median of the distributions. For both indicators and both units the credible intervals conveying the information on the range of the values covering the central 95 per cent of the posterior probability distribution show that there is substantial uncertainty in these indicator results.

Table 5.5: Simulation exercise 2. Observed citation count error-affected and simulated error-free bibliometric indicator values

publication set	observed error-affected values			simulated error-free values		
	P	C	MNCS	P	C	MNCS
A	31	76	0.55	31 (31, 31)	88 (80, 103)	0.58 (0.53, 0.67)
B	38	185	1.07	38 (38, 38)	204 (192, 222)	1.06 (1, 1.14)

Note:

Simulated values are distribution medians and 95% credible intervals in parentheses

Simulation exercise 3 In this exercise, the observed citation counts are assumed to be error-free while the document types are taken as being affected by data inaccuracies. Simulated error-free document types are drawn as predictions from the regression model estimated on the basis of actual document type error data from Donner (2017). Again, 2000 draws from the Bayesian posterior predictive distribution are realized for each input datum. Different total

citation counts and MNCS values can occur even though individual papers' citation counts are not altered in this scenario because different publications can be selected as relevant articles and reviews depending on the predicted document type. The results are shown Table 5.6, where it can be seen that there is some uncertainty in publication counts, total citations, and consequently MNCS. For indicator C in this scenario simulated error-free values smaller than those for the observed data are also predicted. This is because in some cases highly cited documents of an included document type are predicted to be an excluded type and do not contribute to that score. Nonetheless, the median indicator values across all repetitions are close to their observed values and the uncertainties are of moderate magnitude.

Table 5.6: Simulation exercise 3. Observed document type error-affected and simulated error-free bibliometric indicator values

publication set	observed error-affected values			simulated error-free values		
	P	C	MNCS	P	C	MNCS
A	31	76	0.55	32 (30, 34)	78 (71, 85)	0.57 (0.53, 0.61)
B	38	185	1.07	39 (37, 42)	185 (168, 189)	1.07 (1.01, 1.14)

Note:

Simulated values are distribution medians and 95% credible intervals in parentheses

Table 5.7: Simulation exercise 4. Observed error-affected and simulated error-free bibliometric indicator values

publication set	observed error-affected values			simulated error-free values		
	P	C	MNCS	P	C	MNCS
A	31	76	0.55	32 (30, 34)	91 (79, 108)	0.59 (0.53, 0.68)
B	38	185	1.07	39 (37, 41)	204 (184, 223)	1.07 (0.98, 1.17)

Note:

Simulated values are distribution medians and 95% credible intervals in parentheses

Simulation exercise 4 As the final exercise in this series with the same initial observed data we specify that both document type and citation count are affected by realistic levels of inaccuracy. To this end, predicted values for both kinds of data are drawn from the respective regression models as specified above. Results for 2000 simulation runs are displayed in Table 5.7. The simulation

shows that in direct consequence of the empirical data and model structure the observed citation sums are on average underestimated, but also citation sums smaller than the observed ones are consistent with the error distributions. This is due the uncertainty in document types and the inclusion of specific document types. While in the observed data there are 38 articles and reviews for unit B, the simulations show that there would likely be between 37 and 41 such publications with 95 per cent posterior probability if we had error-free data. The results further indicate that the mean normalized citation score, absent errors, would be between 0.53 and 0.68 for unit A and between 0.98 and 1.17 for unit B, with 95 per cent probability each.

5.4.3 A real-world example application

Up to this point, only synthetic exemplary data was utilized. The question remains, does any of this matter in real-world scenarios? To find out, in this section real bibliometric data on the level of formal research groups is analysed.

Data for all WoS-indexed publications of 110 chemistry research groups of six universities in one German federal state was collected in the course of a commissioned provision of bibliometric data in support of a state-wide research evaluation in the discipline of chemistry by an expert panel. Universities and research groups are kept unidentified in this analysis. The data set comprises 3818 different publication records. The groups were asked to submit article and review publications from 2008 to 2013 and for those years there are between about 400 and 700 publications per year, and a smaller number of publications in 2007 and in 2014. The groups published between 3 and 280 papers. Several groups have small publication counts as they were established during the observation period and have only been active for a short period. In order to calculate field-normalized citation impact scores, we use CWTS’s algorithmically constructed classification system, as first described in Waltman & van Eck (2012). Specifically, we use the micro-level fields of the 2020 version of the system, which are the outputs of clustering the citation network with the ‘Leiden algorithm’ (Traag et al., 2019). The publications are distributed over 556 out of 4013 micro fields of that system³⁰. There were no micro level field assignments

³⁰It should be mentioned that the choice of classification system, insofar as it is used for calculating normalized citation scores, can be seen as a source of uncertainty for such indicators, albeit one not based on data inaccuracies. Different classification systems have been shown to produce different item-level normalized citation scores (Haunschild et al., 2022; Robinson-

for 79 records and these are therefore excluded from further analysis³¹. For each combination of publication year and micro class which contains a publication from the chemistry groups we also retrieved all other WoS publication records in order to construct complete reference sets, including citation counts up to spring 2020. This larger set includes 863,443 publication records. Next, document types were simplified by designating all types other than article, review, and letter as ‘other’, so that there are four categories.

As all data are obtained from WoS, we assume them to be subject to the empirically found levels of data inaccuracies and we use the statistical models estimated from original WoS data for posterior predictive inference to obtain likely correct versions of the data on document type and complete citation count, i.e. the models of the second kind from the preceding section. The complete citation count is the WoS citation count plus the predicted number of missed citations from the regression model. The simulated correct document types are also predicted based on the observed WoS document types using the estimated regression model. Because of the larger size of the data set we were limited to creating 1000 estimated data sets of predicted error-free values. This figure is large enough to obtain stable and reliable results.

Figure 5.2 shows the distributions of normalized citation scores (observed citations divided by average citations of publications of the same document type category, publication year, and micro-level field) for a sample of publications of the chemistry groups data set. Grey dots are simulated error-corrected values

Garcia et al., 2023; Ruiz-Castillo & Waltman, 2015; Waltman & van Eck, 2019). It would be difficult to argue that any one particular classification system of the many possible alternatives that could reasonably be used is the most suitable one. Furthermore, some algorithmic classification methods have a random component and repeated runs on the same data with the same parameters can lead to more or less different clustering solutions, the Leiden algorithm being one example. Nonetheless, the choice of this classification system for the present study is deliberate. Unlike the vendor-supplied classification system of WoS, the Leiden system is applied at the level of individual publications, rather than entire journals. Papers of one journal can therefore be classified into different micro-level fields (ideally “topics”), which is highly appropriate for all but the most specialized small journals. Moreover, the Leiden system has the practical advantage that each paper is assigned to exactly one micro field. This should not be underestimated. It eliminates calculating scores of publications for each classification class and then aggregating them, which is required when papers are assigned multiple classes. While algorithmically constructed classification systems are currently not able to capture human-conceived topics perfectly (Haunschild et al., 2018; Held et al., 2021; Waltman & van Eck, 2012), recent empirical research has confirmed that publication-level classification systems, including algorithmic ones, are more accurate than journal-level systems (Colliander & Ahlgren, 2019; Klavans & Boyack, 2017; Ruiz-Castillo & Waltman, 2015; Shu et al., 2019).

³¹These excluded records were, by WoS document type: 43 editorial material, 22 meeting abstracts, 4 news items, 3 biographical items, 3 corrections, 2 articles, 1 letter, and 1 reprint. The two article type records contained no references.

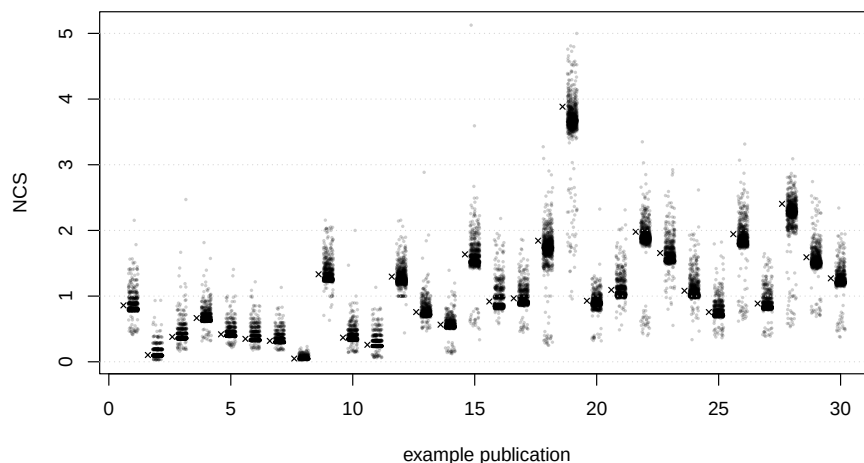


Figure 5.2: Estimates of normalized citations scores for a sample of publications, 1000 runs, symbols: $\times$ original error-affected data, $\bullet$ simulated error-free data

and $\times$ symbols are the uncorrected observed values. While all example items have a central zone in which most of the predictions are concentrated, they also all show substantial item-level variability across Monte Carlo runs. This illustrates the uncertainty propagated from data errors in citation matching and document type assignment to item-level bibliometric indicator scores.

Figure 5.3 shows the uncertainty in bibliometric indicator values on the aggregated level of research groups. The plots show black circles for the medians of the posterior predictive distribution of the indicator values and 95 per cent credible intervals plotted as gray bars for the uncertainty in these estimates, which cover the central 95 per cent of the simulated values. The $\times$ symbols show the uncorrected values. The number of publications typically exhibits relative uncertainty of around 6 per cent on average, computed as the percent ratio of the difference of the bounds of the credible interval to the value of the median, the best estimate for the error-free value. The values of research groups with small numbers of publications are barely affected by uncertainty, as they have very low absolute chance of being affected by document type assignment errors. The largest calculated credible interval is for the group with the greatest number of observed publications. This group reported 257 publications and its 95 per cent

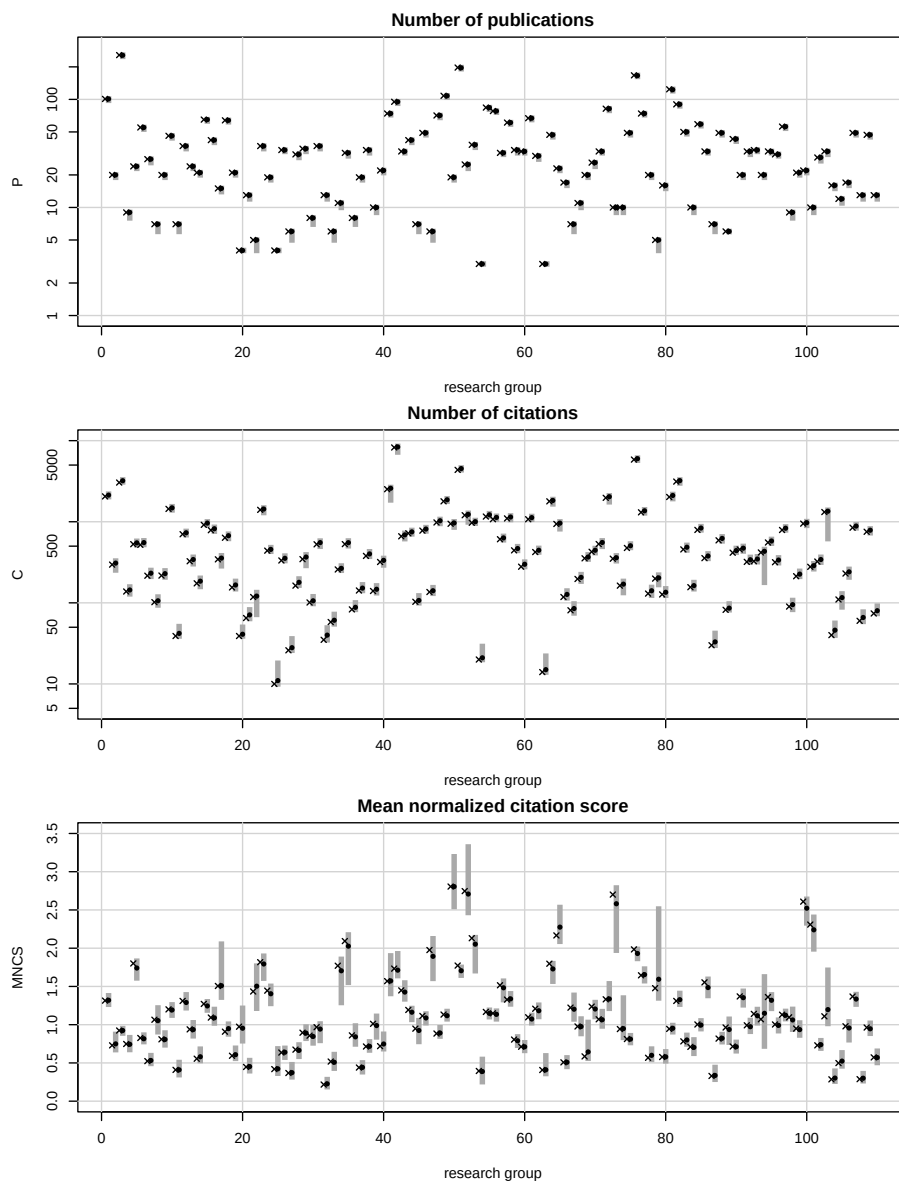


Figure 5.3: Estimated error-corrected data for publications (P), citations (C), and mean normalized citation score (MNCS) for 110 chemistry research groups, medians and 95% credible intervals, 1000 runs, symbols: $\times$ original error-affected data, $\bullet$ simulated error-free data

credible interval is 250 to 257 with a best estimate (median) of 256.

Relative uncertainty in group citation sums is greater, being around 18 per cent on average. Readers should bear in mind that the y-axes on these two plots are logarithmically scaled to better show the whole range of values, which contracts the apparent size of uncertainty bars in the higher values. Uncertainties are quite large for groups in the whole range of total citations. The group with the smallest reported value of 10 citations has a credible interval from 10 to 17 citations, or 70 per cent. On the opposite end of the spectrum, the one with the largest reported value of 8239 citations has a credible interval from 7219 to 8490, or 15 per cent relative uncertainty.

The relative uncertainty in the group level estimates of MNCS is on average 23 per cent. Figure 5.3, bottom panel, clearly shows that higher values are associated with greater uncertainties. The largest uncertainty is found for a group with a best estimate MNCS of 1.6, with credible interval of 1.3 to 2.5, while its uncorrected MNCS value is 1.5 – a 79 per cent relative uncertainty. Best estimates of error-free MNCS values are typically close to the observed error-affected values but associated with large uncertainty intervals.

Figure 5.4 shows the relationship between the best estimate of the publication count, P , and the relative uncertainty in the MNCS. Each data point is one research group. There is a clear and robust association, groups with fewer publications have MNCS values associated with greater uncertainties. However, this relationship is far from perfectly log-linear, as there are a number of groups with much higher uncertainties than others with similar publication counts. Even for groups with hundreds of publications, the relative uncertainty can be on the order of 80 per cent.

5.5 Discussion

5.5.1 Summary

In this study we have outlined a general approach of the estimation of uncertainty in bibliometric indicators. It consists of using empirical data on item-level error distributions for bibliometric quantities to estimate Bayesian regression models and generating predictions from the model for new data as Monte Carlo simulations. This allows the propagation of input uncertainties whose distributions are known to any bibliometric indicators of interest. Thus the uncertainty

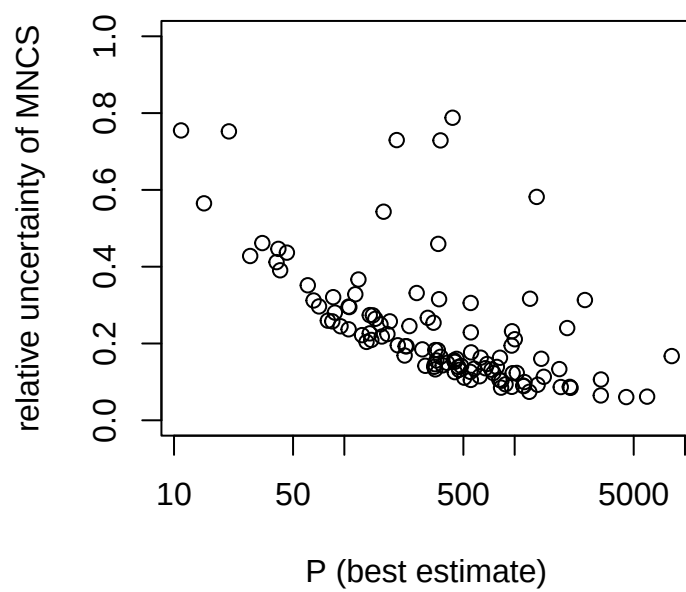


Figure 5.4: Relationship between P and uncertainty in MNCS, 110 chemistry research groups

in bibliometric indicators that is due to data quality issues can be explicitly modelled and expressed. This uncertainty can be expressed, for example, as median estimates of a quantity and corresponding credible intervals corresponding to a chosen probability that the likely correct value lies within that range. We have demonstrated this approach with a number of simple synthetic examples for didactic purposes and for one example of real-world data with a specific evaluative bibliometric application. These examples showed substantial uncertainty in bibliometric indicator values that would otherwise have remained invisible.

5.5.2 Limitations and future work

The present study has several limitations. We have worked only with one moderately sized manually assessed data set for the validation of citation counts from a single data source. This is clearly not generalizable to every scenario but can at best serve as one example to be extended and supplemented with further samples in the future. Our results do not transfer to citation counts from any other data source or even the same data source once the proprietary algorithm for reference matching is changed substantially. As a case in point, the study by Van Eck & Waltman (2017) shows that the types of errors in citation data differ by data source. Furthermore, we have used a single predictor variable, WoS observed citation counts, to predict omitted citations. This implies the assumption that the number of omitted citations is otherwise not affected by any characteristics of the specific publication. This is a necessary simplification due to constraints in data collection. There are other relevant predictor variables for omitted citations. One example would be the scientific field of a publication. For instance, in the humanities, complex footnote citations are common, as opposed to reference lists. Such citation data is inherently more difficult to extract correctly and completely, which could manifest in greater citation error in this domain, as results in Tüür-Fröhlich (2016) suggest. Another example for a possibly useful predictor variable would be publication year. Publishing practices and data collection practices of citation index databases change over time. Therefore data originating in different time periods might be differently affected by citation count errors. A third example would be publications from languages other than English being cited as references. These may possibly exhibit greater error rates than English-language references if data extraction methods are primarily designed for English language data and perform worse for other languages. In many cases such cited references are also cited in trans-

lated or transcribed form, which are practices susceptible to the introduction of errors or variability, such as when competing romanization systems are used for non-Latin scripts. Moreover, the research of Moed & Vriens (1989) and Buchanan (2006) showed that specific journals are more affected by referencing errors, as are papers by authors with complex names. Including such other predictors as covariates into regression models like the ones discussed here would require larger samples sizes in order to achieve reliable estimates across ranges of numerical and categories of categorical variables.

Another limitation is the choice of specification for the regression models used in the stochastic prediction of correct and error-affected data. For the citation error model, we chose a negative binomial model as it is a count data model well suited to skew data such as citation counts. Other choices are possible and perhaps better. The purpose of the present study was not to identify the single best specification choice, for that the size of the sample was not large enough, but rather to outline one approach and test it for one reasonable model choice. Moreover, incorrect document types and incomplete citation counts are just two of a larger number of known data error-related phenomena. We have focused on omitted citations, a source of uncertainty which is documented and understood relatively well and is assumed to have considerable effects on data quality. Thus we have not taken into account other phenomena of citation error such as phantom citations or unindexed citations from supplemental material. Ultimately all sources of data error need to be either eliminated by correction or included in a more comprehensive error assessment of bibliometric indicator results. We expect that increasing the sample sizes, including additional predictors and using an alternative regression model specification would lead only to minor quantitative changes in the results we report here. On the other hand, the inclusion of more types of data errors can be expected to lead to increases in predicted uncertainties. In this sense, the present limitations mean that our reported uncertainty intervals are probably conservatively small.

We have restricted the study to three very basic bibliometric indicators, publication counts, citation counts, and mean normalized citation scores. While most bibliometric studies use the first two indicators, the MNCS indicator is just one of the many possible choices to characterize the typical citation impact of a set of publications. It is known that this indicator can be sensitive to influence by a few very highly cited papers. A more comprehensive study of the influence of data errors on other important impact indicators, including representatives of

other classes of citation impact indicators, such as highly cited rate indicators and indicators of total impact, is certainly warranted. It might be prudent to investigate if any citation impact indicators are particularly volatile or robust to data errors.

This study only looked at one kind of research unit when studying real-world data, formal university research groups. Again, this example was chosen for being illustrative. It is arguably the smallest type of unit that is reasonably amenable to bibliometric scrutiny, as individual researchers do not commonly have the necessary number of publications to make statistical analysis reliable. This is not even strictly the case for the research groups used here. The intention behind the choice of this small a unit was to make any possibly found uncertainties easy to perceive. It remains to be studied if the quite substantial uncertainties found here can also be seen in larger units, such as departments or whole organizations, or if they might become small for large enough publication sets. In any case, the uncertainties at the level of single publications will not disappear until the data are free of any error.

5.5.3 Conclusion

In the first empirical part of this work we studied the incidence of missing citation links in WoS. According to a global random sample, about 4 per cent of citation links are missing in WoS and nearly one third of all studied items had one or more missing citations. The citation error distribution was found to be highly skewed. It should be pointed out that citation index providers continue to improve their databases including their original data indexing quality and their reference matching procedures, so our results represent a snapshot of data quality at one particular point in time. Furthermore, both WoS and Scopus welcome corrections and incorporate them into their data after confirmation. So any discovered errors or omissions could in principle be corrected in the data source.

In the second empirical part we used these data and data from an earlier study to statistically model the errors in citation counts and document type assignments. We demonstrated how such models can be used to simulate error-corrected data from typical error-affected real data to gain insight into the range of bibliometric indicator values consistent with the data at hand and known error distributions. This method has the potential to improve the reporting of bibliometric indica-

tor values and thus the interpretation of the results of bibliometric studies. It provides one practical implementation of the eighth principle of the Leiden Manifesto, which states that information on uncertainty and error in values should be quantified and reported alongside indicator values whenever possible.

By studying real bibliometric data that was used in assessment of university research groups in chemistry, we were able to show that data errors do not just cancel out and vanish at this level of aggregation but lead to quite substantial uncertainties in the values of bibliometric indicators. The hitherto indifferent attitude of professional bibliometricians to data quality and data errors can hardly be sustained in the light of the results presented herein. Uncertainty quantification can contribute to the legitimacy of the field of scientometrics because it increases transparency about known inaccuracies in data, instead of ignoring them.

Acknowledgements

I want to thank Nees Jan van Eck for in-depth discussion of the manuscript and for providing the algorithmic classification data. I also thank Stephan Stahlschmidt for helpful feedback on the manuscript, Anastasiia Tcypina for help in collecting the empirical data, and Sybille Hinze and Stefan Hornbostel for first introducing me to the concept of ‘calculus of errors’ for bibliometrics. I am also grateful for the thoughtful suggestions of two anonymous referees.

Appendix

5.5.4 Simulating error-affected data

In this section, we demonstrate the use of simulation of realistic errors for given error-free data sets. These are the models of the first kind of the section *Incorporating information about uncertainty into statistical models*. They are probably of more limited practical use but very instructive nevertheless. For reasons of exposition, we will begin with a small example data set and consider the two sources of data error, omitted citations and inaccurate document types, first in isolation and then in combination.

The number of missed citations, a count variable, is modelled using a negative binomial regression model with a single predictor, the accurate number of citations. This predictor is modified by adding 1 and taking the natural logarithm:

$o_i \sim \text{NegB}(\ln(c_i^* + 1), \theta)$. This notation also follows Franceschini et al. (2013): o_i is the number of omitted citations, c_i^* is the accurate number of citations; θ is the negative binomial overdispersion parameter, estimated from the data.

The simulated error-affected citation count can then be obtained by subtracting the predictions from the model for new data from the input values. A drawback of the particular model is that it can predict numbers of missed citations greater than the number of true citations. This would result in predicted negative citation counts which is incongruous with reality. We therefore set a lower limit of 0 on the simulated citation counts: $c_i^{\text{sim}} = c_i^{\text{new}} - \hat{o}_i$ subject to: $c_i^{\text{sim}} \geq 0$.

Simulation exercise A1 We generate a synthetic data set of known accurate values for two research units, A and B, with 20 and 30 publications, respectively, of various document types. For calculating bibliometric indicators, only articles and reviews will be selected and counted (indicator publication count, P) while letters and items of other document types are discarded, as is common in bibliometric studies. The units have only published in one homogeneous research discipline and in the same publication year, so no further normalization for such factors needs to be applied. An additional reference set of publications consists of 1000 publications of all document types. Document types of all three publication sets are sampled from the global document type distribution which, based on WoS data, is as follows: articles: 68 per cent, letters: 3 per cent, reviews: 4 per cent, other: 25 per cent. Citation counts of papers, the input quantity in this exercise, are generated from discretized lognormal distributions such that units A and B and the other publications each have their own typical citation impact levels, realized by setting the mean parameters of the distributions the citation count values are drawn from at different values (A: ln 5, B: ln 7, ref.: ln 6). Depending on the document type these parameters are scaled with multiplication factors as follows, with articles at 1.0: reviews: 1.5, letters: 0.2, other: 0.1. For the two units, the three basic bibliometric indicators number of article and review publications (P), number of citations of these (C), and mean normalized citation score (MNCS), are calculated. The item-level normalized citation score (NCS) of a publication is calculated as the observed citation count divided by the expected citation count. The expected citation count is the average of the citation counts of all publications of the same document type. The MNCS of a unit is the arithmetic average of the NCS of its publications (Waltman et al., 2011). The choice of indicators is purely illustrative and the present approach

can be used for any other indicators.

The generated synthetic correct values are used as inputs to the statistical error model of the first kind for citation counts, which was estimated using the empirical data from the first part of the paper. In this exercise, no error is assumed for document types, so that data is unchanged. From the model, 2000 samples of the posterior predictive distribution of the omitted citation count for each publication item are drawn. In each case, the simulated error-affected citation count of a paper is calculated by subtracting the predicted number of omitted citations from the error-free original citation count. For each such set of simulated error-affected data the bibliometric indicators (P, C, MNCS) of the two units are calculated. By this method we can obtain distributions of the indicator quantities of interest generated from Bayesian posterior predictive distribution samples, given the known error-free data and the statistical model estimated using empirical data. In other words, distributions of the probable data that might have been available from WoS given the knowledge of the error distributions in this data source and the values of the input quantities.

The correct values and summaries of the obtained distributions of simulated error-affected values for the bibliometric indicators are reported in Table 5.8. There is no variability due to document assignment error, so the document types and thus the values for P are constant. Citation counts in the simulated inaccurate data are lower than the correct figures. Median simulated MNCS estimates are not far off their error-free values but the uncertainty is substantial. Appendix Figure 5.5 shows the results for 2000 repetitions.

Table 5.8: Simulation exercise A1. Error-free and simulated citation count error-affected bibliometric indicator values

publication set	error-free values			simulated citation count error-affected values		
	P	C	MNCS	P	C	MNCS
A	16	122	0.76	16 (16, 16)	116 (102.975, 121)	0.77 (0.68, 0.8)
B	20	238	1.14	20 (20, 20)	225 (208, 234)	1.15 (1.06, 1.2)

Note:

Simulated values are distribution medians and 95% credible intervals in parentheses

Simulation exercise A2 In this exercise, we look at the isolated effect of error-affected document type data on bibliometric indicator uncertainty. Citation counts are not changed from their original error-free values.

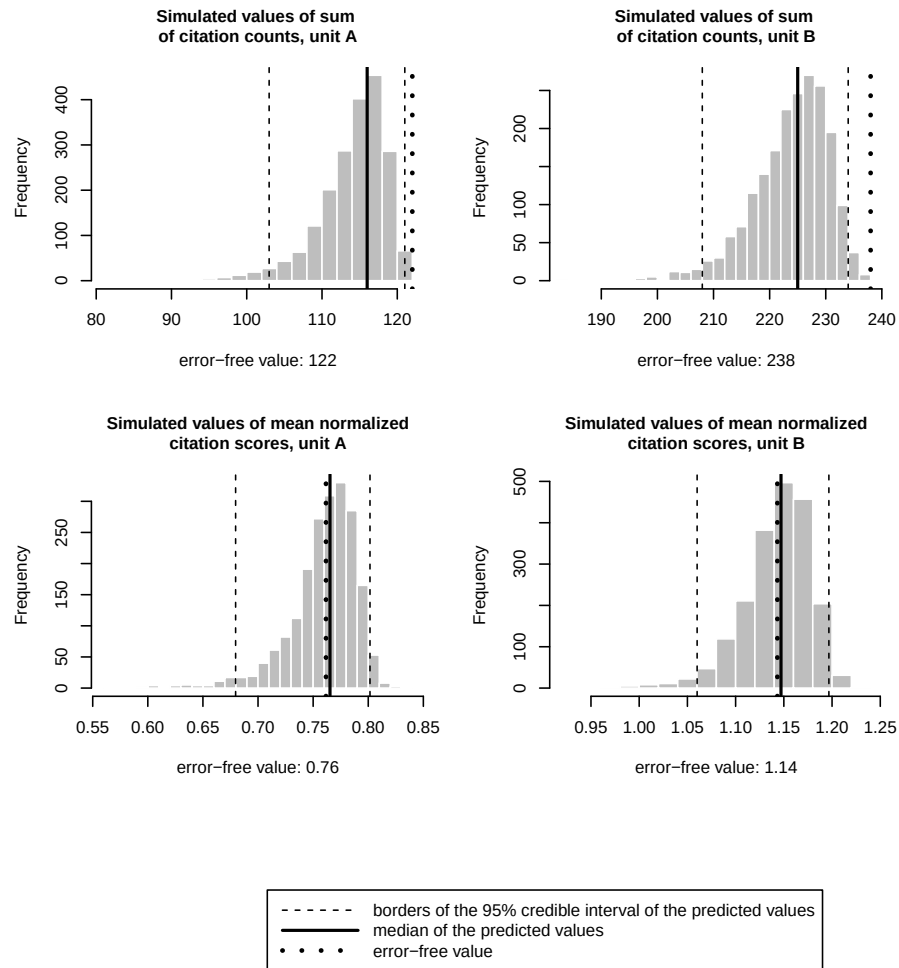


Figure 5.5: Results simulation exercise A1 - data affected by citation count error

For the model of error-affected document types, we create a categorical distribution regression model that predicts the possibly error-affected document type d_i (four classes) assigned by the database vendor from a single predictor, the known correct document type d_i^* (same four classes): $d_i \sim \text{categorical}(d_i^*)$. The model is estimated with the empirical data on document type inaccuracy in WoS from Donner (2017). The estimated model is used to run a simulation to illustrate the influence of realistic document type error using the setting of the prior exercise. In contrast to that exercise, here the citation counts of items are not affected by any error and thus constant. Again, 2000 sets of values are simulated and the three basic indicators calculated. The results are shown in Appendix Table 5.9. It can be seen that all aggregate level indicators are affected by document type errors: publication count because different numbers of documents with the document types of interest are counted and citation count for the same reason. MNCS is affected in addition to that because of distorted reference group citation count values of document types. We can see that indicator value uncertainties in this exercise are greater than in the previous one in which only citation counts were affected by errors.

Table 5.9: Simulation exercise A2. Error-free and simulated document type error-affected bibliometric indicator values

publication set	error-free values			simulated document type error-affected values		
	P	C	MNCS	P	C	MNCS
A	16	122	0.76	12 (8, 15)	105 (42, 119)	0.84 (0.37, 1.13)
B	20	238	1.14	15 (11, 19)	185 (122, 231)	1.16 (0.83, 1.45)

Note:

Simulated values are distribution medians and 95% credible intervals in parentheses

Simulation exercise A3 What happens if both types of errors are present? This can be studied by simulating error-affected values from both the document type error model and the citation count error model. Using the same input data as before, the results are shown in Appendix Table 5.10. The results are comparable in uncertainty to those from the simulation exercise in which only document type error was modelled. In this particular scenario the contribution of document type error to overall uncertainty in bibliometric indicators is clearly greater than that of citation count error. Such a finding could only be uncovered by this modelling-prediction-simulation approach.

Table 5.10: Simulation exercise A3. Error-free and simulated citation count error and document type error-affected bibliometric indicator values

publication set	error-free values			simulated citation count error- and document type error-affected values		
	P	C	MNCS	P	C	MNCS
A	16	122	0.76	12 (8, 15)	98 (38, 115)	0.84 (0.36, 1.15)
B	20	238	1.14	15 (12, 19)	176 (115, 222)	1.15 (0.83, 1.45)

Note:

Simulated values are distribution medians and 95% credible intervals in parentheses

Simulation exercise A4 The first set of three scenarios had units with deliberately small publication counts to support conveying the basic approach. At such low sample sizes the effects of errors are easy to observe from the large variability they introduce. However, this variability could conceivably be a minor factor in the much larger sample sizes which are common in bibliometrics. To test this hypothesis, we now investigate what happens if the publication counts are much higher. We keep all settings as before but increase the number of simulated publications for all three publication sets, A, B, and ref., by a factor of 100 to 2000, 3000, and 10,000 publications, respectively. The results are reported in Appendix Table 5.11. The accuracy of the simulated error-affected MNCS indicator value has now improved a lot, yet there is still an appreciable uncertainty. Moreover, the possible error-affected values for P and C exhibit a wide range of results in absolute terms and are not close to their error-free values.

Table 5.11: Simulation exercise A4. Error-free and simulated error-affected bibliometric indicator values

publication set	error-free values			simulated error-affected values		
	P	C	MNCS	P	C	MNCS
A	1472	12813	0.84	1090 (1005, 1169)	9009 (8273, 9725)	0.84 (0.8, 0.87)
B	2202	26977	1.19	1628 (1504, 1734)	19057 (17553, 20268)	1.19 (1.15, 1.22)

Note:

Simulated values are distribution medians and 95 per cent credible intervals in parentheses

These results show that simulating realistic errors with different conditions can be informative about the contribution of different types of errors to overall indicator value uncertainties and about how the accuracy of indicator results changes with sample size.

6 General conclusion

The preceding series of four empirical studies was designed to generate new insights on the presence and severity of inaccuracies or errors in bibliometric data and their consequences. They contributed to answering the general research question of the dissertation, which is: *How does data quality affect the reliability of bibliometric indicators?* A particular focus was on the data quality dimension of accuracy as a determinant of uncertainty in bibliometric indicators. Beyond this, the studies addressed also the particular gap in the literature of bibliometrics relating to the quantification of inaccuracy and uncertainty for common bibliometric indicators which originates in data quality. In doing so, they have improved our understanding of the relationship between errors in data and uncertainty in indicator values.

6.1 Summary of findings

The preceding studies have made one thing abundantly clear: while bibliometrics researchers may commonly assert the objectivity, numericity, and accuracy of their methods, if we lift the carpet we find that the underlying basic bibliometric data are laced with the dust of errors and inaccuracies. These inaccuracies, far from vanishing and canceling each other out in large scale analyses, affect bibliometric indicator quantities to a degree that cannot be ignored. This is something that many experienced bibliometricians may have vaguely anticipated. It is the “false precision” of the Leiden Manifesto which ought to be avoided (Hicks et al., 2015). The studies have shown how selected data quality issues are responsible for considerable uncertainty in bibliometric indicators from the level of single publications’ normalized citation scores to the level of aggregate scores of output and impact of research groups and research organizations.

The overall common approach that I used was to assemble original data from bibliometric information systems, as they are used in standard bibliometric studies, and contrast this with either manually or automatically corrected versions of the same data. The observed differences were then statistically investigated to derive error rates and characterize error distributions in some detail. Next, using both the error-affected and the corrected data, I looked at how much the found errors affect typical bibliometric studies and indicators to assess the real-

world importance of the findings. This is an essential task because just knowing that the rate of errors in some bibliographic field is 5 or 10 % by itself does not tell us if the inaccuracies cause any serious problems or can be ignored because bibliometric indicator result values are not affected to any degree that matters. Finally, the approach culminated in a proposed framework for quantifying the uncertainty in indicator values by statistical models that are constructed from the empirically collected error data.

The study of the accuracy and reliability of bibliometric databases' assignments of document type information to records in Chapter 2 found that this information is affected by errors with considerable frequency. For the WoS sample of documents, I found that the error rate for document type assignments is about 6 %. I also found that the document type categories assigned by different databases, in this case Web of Science and Scopus, to the same publications, are not very consistent, even though the definitions of categories are comparable. The category assignments for three major document types and a residual 'other' category for a sample of 722 publications were consistent between databases in 83 % of cases. The influence of these errors on citation scores was also studied. Inaccurate document types of a seemingly small fraction of documents do not only affect citation scores of these incorrectly classified records themselves. In any study which uses normalization of citation counts by document type or selection of records by document type, they affect the normalized citation scores of all records. The reasons for this are that (1) they cause the calculation of inaccurate expected citation scores and (2) lead to the inadvertent inclusion or exclusion of the wrong records from analyses. These findings are complemented by another, more unexpected result. I found that a small share of publications cannot be unambiguously classified into exactly one discrete class of document type because they have characteristics of two classes. The findings of Chapter 2 on document type classification error have already been independently substantiated. A similar study was carried out for a more specific sample of highly cited publications by Russian authors (Mokhnacheva, 2023) which confirmed the main findings of my study. The findings were further corroborated for a sample of Italian publications by Maisano et al. (2025), who also exploited the observed between-database discrepancies to semi-automatically identify possible document type errors and verify them at a much larger scale. The study also motivated researchers at the Chinese Academy of Sciences to check their revised journal impact indicator for improved robustness to document type error (Tong

et al., 2023).

The study on publication months (Chapter 3) has shown that publications which appear earlier in a given year have a citation count advantage over those that appear later in the same year when using conventional year-based citation windows. I argued that the existence of this publication month effect means that conventional year-accurate citation window methodology is biased. I showed that month bias persists even for citation windows of over five years. This is a systematic error and hence, with sufficient data on publication month, can theoretically be eliminated. Publication month data are not always present in citation databases. I demonstrated that publication month can reliably be inferred from journals' issue order information. I showed further that when using publication month data to construct equally long citation windows for papers published in different months of a year, this publication month bias is successfully eliminated, as theoretically expected. This means that month-precision citation counts have higher validity than year-precision citation counts for any latent construct for which they are invoked as indicators. One may wonder if this matters in practice. Bibliometrics is primarily used to statistically analyze larger publication sets, such as all publications of some research performing unit in a discipline in a certain year. When such sets of publications are studied as aggregates, it seems intuitive that there will be a mix of early-months and late-months publications and the bias components of citation scores will cancel out. To test this intuition, I carried out a realistic citation analysis of German research organizations at the level of OECD Fields of Science disciplinary categories and compared average normalized citation scores computed either with (month-biased) nominal 5-year citation windows or actual 5-year (60 months) citation windows. I found that this improvement in citation score accuracy leads to clearly discernible score differences which are in many cases of substantial magnitude. It has to be concluded that publication month is also in practice a nuisance variable which introduces bias in citation analysis, even for analysis of larger numbers of publications.³²

The comprehensive investigation of the data quality of the institutional disambiguation systems of mainstream commercial bibliometric databases in Chapter 4 has shown that these systems are very incomplete in their coverage of

³²From the regression results presented in Table 3.7 it can be deduced that publication month accounts for 2 to 4 % of variance in 3-year logarithmized citation counts of individual papers.

organizations and have in many cases poor precision and recall in the identification of papers that should be attributed to organizations. Many organizations are not covered as disambiguated entities in the platform systems of WoS and Scopus at all, although their papers can be found by organization name searches. My co-authors and I found that the WoS and Scopus systems, for the organizations they cover, on average exhibit adequate precision but worse recall. However, precision and recall for organizations within one system varied greatly such that results for some organizations were much worse than for others. We further found that bibliometric indicator values calculated from the vendor databases' disambiguation systems have only very limited accuracy compared to reference values calculated from German organizations' data disambiguated with a known very high quality. The conclusion from this study is that we recommend additional data processing for any use with policy implications. In summary, the platform disambiguation systems were shown to have insufficient accuracy in elementary bibliometric indicators such as publication counts of organizations to be reliable enough for professional bibliometric use. The main results of this study have been independently corroborated by Selivanova et al. (2019) and Dederke et al. (2024) for other countries.

Chapter 5 established, for the first time, that the error distribution of citation counts (specifically missing citations) due to data quality deficiencies (incomplete citation matching) does not follow a normal distribution, as is common for error distributions in the sciences and engineering. Most papers have no or very few missing citations but a small number of papers have high numbers of missing citations. This means that errors, that is to say, missing citations, are closely clustered into relatively few papers which are highly affected and whose actual citation counts are severely underestimated by the observable citations. In other words, the error distribution is skewed. This is a fundamental empirical finding which links the error distribution firmly to the skewed distribution of citation counts and of most other bibliometric quantities. Chapter 5 also confirmed a number of other studies that worked with non-probability samples by showing for a proper random sample the rate of omitted citations in WoS to be on the order of 4 %. Chapter 5 has gone much beyond the mainly empirical-comparative research of the other parts of the dissertation by proposing, introducing and demonstrating a novel framework for the formal propagation of uncertainty from known inaccuracies, such as data quality problems, via measurement models to final bibliometric indicator values. This framework is

general in that it is applicable to any source and any kind of data inaccuracy that can be empirically studied and quantified, and to any bibliometric indicator. Using this new framework and applying it to a realistic bibliometric research assessment scenario, I found that the bibliometric data quality uncertainty from just two sources of error, missing citations and inaccurate document types, is considerable, both on the level of the normalized citation impact of individual publications, but also on the counts and aggregate citation impact scores of the publication sets of research groups.

Incidentally, as an error rate, can we say anything more, qualitatively, about the magnitude of the figure of about 4 % missing citations? Is it a low or a high value in its context? There is, of course, no objective yardstick to compare this to. But I want to offer the following isolated scalar value as a point of comparison. Four hundred years ago, English mathematician Henry Briggs worked out huge logarithm tables by hand. The error rate of this work is surprisingly low:

Briggs calculated each value up to the thirtieth decimal place, and he wrote them down following Stevin's then still novel notation, but already with a period separating the integer from the fractional part (a practice that Napier had initiated). With the help of the rules of multiplication and division of logarithms, Briggs constructed a table with very dense values. What he was able to calculate in his lifetime he published by 1624. This comprised logarithms of all numbers between 1 and 20,000 and then between 90,000 and 100,000, which he wrote as decimal fractions with 14 digits after the period. Throughout the years, about 1100 mistakes were found in the tables. This amounts to no more than 0.04% of the total entries. In most cases, these were deviations of ± 1 in the last digit, plus some simple typos. (Corry, 2015, pp. 169–170)

To put that into perspective, Briggs' logarithms had an error rate two orders of magnitude lower than modern citation counts. This makes the citation matching error rate appear subjectively high.

To return to the finding of a skewed distribution of missing citations, one important implication of this should be pointed out. The seemingly moderate rate of missing citations of 4 % is an average – an average of a skewed distribution. This means that it is not the case that typically every paper will have 4 % of

its citations being missing. Rather, most papers have no or a negligible number of missing citations. But a small number of papers will have a great number of missing citations.

Having reviewed the main results of the studies, I am now in a position to answer the general guiding research question of the doctoral studies project: *How does data quality affect the reliability of bibliometric indicators?*

In the cases of the error rates for document type assignments and omitted reference links, the empirical estimates may appear to be of small magnitude. But the complementing example studies of realistic research assessment analyses have shown that this impression is misleading. These small error rates cause substantial uncertainty in bibliometric indicator values. The case is more straightforward for the organization disambiguation systems of leading proprietary bibliometric databases. These were found to exhibit such poor performance that their use for any serious research or applied studies was cautioned against.

The overall answer to the main research question is therefore that data quality affects the reliability of bibliometric indicators negatively to such a degree that standard bibliometric indicators are usually very imprecise. They are imprecise in the sense that they are typically not very close to the theoretical value that could be calculated if all data errors were corrected. The broad implication is that bibliometric indicator values should routinely be treated as inaccurate estimates with unknown, but substantial, uncertainty. More implications and recommendations follow below.

6.2 Consequences for bibliometric studies and bibliometric data sources

6.2.1 General implications

The studies of this dissertation have demonstrated unambiguously that indicator values calculated from data by commercial data providers using accepted standard methods are affected by substantial levels of uncertainty. Like any measurement, they come with a certain level of unreliability. The cause of unreliability studied in this thesis is insufficient data quality. This core finding has considerable implications that all providers and users of bibliometric

data, indicators, and reports should be aware of and take seriously. Precisely reported numerical values of indicators should be interpreted as estimates with unknown, unspecified uncertainty. Unless uncertainty estimates are actually present, which would be the much preferable option. My studies into the real-life effects of indicator uncertainty have shown that indicator value differences on the order of perhaps 20 to 30 % for research units with hundreds to thousands of publications, such as the multi-year output of research groups, research institutes, or universities, cannot be taken at face value. Such differences could be caused simply by inaccurate base data.

Bibliometric indicators and measurement methods should be continuously reviewed, modified and improved (Hicks et al., 2015). One methodological improvement whose broad implementation I want to advocate for is to substitute month-precision citation windows for year-precision citation windows to remove month bias in citation indicators (Chapter 3). This change is easy to implement, both for database users and for database providers. It can measurably improve the reliability, and hence also the validity, of citation counts, irrespective of the construct for which they are invoked as a proxy.

6.2.2 Research assessment reform

The debate around reform of research assessment practice and the roles of qualitative and quantitative evidence has increased in scope and activity in recent years. Much of this debate has been implicitly about the level of individual researcher assessment, where a common position of advocates of bibliometric methods is that these methods can be a helpful complimentary source of evidence. At this level, their arguments revolve around the low inter-rater reliability of peer review procedures and their susceptibility to particularistic social biases.

My studies have demonstrated clearly that bibliometrics has its own reliability problem. For example, at the publication level, citation counts are frequently inaccurate and document type assignments exhibit significant error rates. The latter issue means that when publications of selected document types are counted, for example the publication of type ‘Article’ of one researcher, these counts are also affected by inaccuracies. I have not studied the researcher aggregation level, but the next higher level, that of formal research groups. The errors from the paper level have been propagated to the research group level indicator

values and due to their non-normal distributions, they do not appreciably statistically cancel out. The indicator value uncertainty levels strongly affect the interpretation of the results. Under the same or similar error levels, uncertainty ranges for indicator values on the individual researcher level can be expected to be larger in relative terms because of the smaller number of observations at this level. Therefore, the use of bibliometrics in researcher-level assessments with unchecked and uncorrected data from major bibliometric data providers at present data quality levels must be cautioned against.

There are two obvious ways of metrics-advocates to respond to these findings. The first is to substantially improve data quality. The the researcher level, this is to some extent possible. One can let researchers check publication lists for completeness and correctness of document type and one can correct citation counts by systematic cited reference searchers. Similar steps would need to be taken for other data quality issues. That would make the base data as accurate and high-quality as it reasonably can be made. The second way would be to explicitly report the proposed uncertainty ranges for indicator values that derive from data quality and error propagation. However, simply pretending that unreliability is a problem only for qualitative assessment of research and not for bibliometrics can no longer be justified.

6.2.3 Improvements in data quality

From studies such as these presented here, which show that poor data quality seriously thwarts reliable and accurate bibliometric studies and from the emerging competition of open databases, one might hope that commercial providers decide to invest in improving data quality to advance the competitiveness of their products. It would be rational to do so because higher data quality means more reliable bibliometric scores which can serve as a basis for more legitimate and evidence-based decision-making. The continued use of data of too low quality may eventually become a liability if these inadequacies are taken up as an argument against specific uses of bibliometric data and methods with relevance for policy or personnel decision-making. It seems likely that commercial database providers have so far not fully appreciated the possible adverse influence of insufficient data quality on the accuracy of results calculated from their data by users and by themselves. Thus, it may be a worthwhile longer term strategy to implement further data cleaning routines or revise existing ones.

In Chapter 4, my co-authors and I found that research organizations and the research publications that they have produced are often difficult to identify comprehensively and accurately in bibliometric databases. The recommendation that follows from these findings is that the on-board institutional affiliation disambiguation systems of Scopus and Web of Science should not be used without further processing and cleaning until the retrieval performance of these systems has improved. The operators of these databases should continue to improve their systems and publish their internal quality assessments to substantiate claims about their performance.

One aspect of the institution identification problem has been partly addressed by the introduction of the Research Organization Registry (ROR). ROR is a “global, community-led registry of open persistent identifiers for research organizations”,³³ which at the time of writing comprised more than 116,000 organizations. ROR’s data is openly available. ROR solves the issue of the usage of different identifiers across different bibliometric and scholarly information systems as its identifiers can be used without cost and licencing restrictions. OpenAlex and Web of Science have adopted ROR IDs.

6.2.4 Open data, open methodology, and transparency about uncertainty

The fundamental possibility to do any data quality studies of bibliometric data sources was until recently very restricted. Access to comprehensive raw data was only available at a few dedicated research centers because of the high costs and restrictive licenses of commercially produced data sources. These were the conditions under which this doctoral research project was started. Meanwhile, conditions have profoundly changed. Large-scale, comprehensive, open bibliometric data sources have been launched; sources with no licencing costs and permissive use licencing. Consequently, studies of the quality and reliability of bibliometric data are now within reach of many more researchers and users. Data quality studies of completely different scale, range, and openness are now possible with open data.

The availability of open bibliometric data has enabled consultants and contractors as providers of bibliometric reporting services to meet the natural and long-standing, but legally difficult to arrange expectation of transparency of

³³<https://ror.org/>

the underlying data to the evaluated units,³⁴ thereby contributing to accountability and mutual trust. A knock-on effect of open data is that bibliometric methodology can also become more transparent. Bibliometric research groups were hesitant to make their technical methodology openly available for several reasons. One reason was that their technical infrastructure, such as processing routines, were so closely bound up with proprietary data that the disclosure of source code may have violated licensing terms.

A prominent example of a transformation from proprietary to open data in bibliometrics is the Leiden Ranking Open Edition³⁵ which uses OpenAlex data and fully open source code³⁶. Another example of improved transparency in bibliometric research methodology is that the KB affiliation disambiguation system, used extensively in Chapter 4, has been open sourced as one early step of a larger movement of the German bibliometrics and science studies community towards open science practices, reproducibility and transparency (Lenke & Taubert, 2024).

This development towards open bibliometric data may seem to somewhat diminish the long-term importance of completed data quality studies of proprietary bibliometric databases. However, studies of proprietary data sources such as these presented here, retain their usefulness in that their results can serve as points of reference for the comparison of the data quality of open bibliometric data. Given the not very high data quality observed in the empirical analyses in this project, the expectation for open bibliometric data sources should be to eventually surpass the data quality of commercial products. The open availability of bibliometric data is a great advantage but this advantage can only be fully realized and utilized if the quality of the data is sufficiently high to meet the quality standards of the research community and of the community of decision-making users of bibliometric reporting products. Consequently, at this moment, there is a lot of research on the data quality of the foremost open bibliometric database product, OpenAlex, with presently a strong emphasis on data completeness (Céspedes et al., 2025; Culbert et al., 2025; Haupka et al., 2024; Hauschke & Nazarovets, 2025; Rodrigues et al., 2025; Scheidsteger et al., 2025; Zhang et al., 2024). This shift of attention of empirical data quality studies from proprietary to open data sources seems to be not only a worthwhile

³⁴San Francisco Declaration on Research Assessment <https://sfdora.org/read/>

³⁵<https://open.leidenranking.com/>

³⁶<https://github.com/CWTSLeiden/CWTS-Leiden-Ranking-Open-Edition>

investment of the research community's effort but also an early indication of an imminent change in the usage patterns of data sources of the scientometrics and science studies community.

A crucial development precipitated by open bibliometric data is the possibility of transparency of bibliometric studies, including both full reproducibility of research work and full disclosure of basic data and calculations for bibliometric research assessments. With this possibility comes of course the expectation that such transparency is the default mode. Heretofore, licensing terms of commercial bibliometric data subscriptions have precluded such transparency. This condition may have contributed to concerns about the legitimacy of bibliometric evaluations since fundamental data could not be independently verified. Clearly, transparency is highly desirable. I believe that studies of data quality, reliability, and validity can contribute further to transparency. Such studies can complement basic data transparency by demonstrating the uncertainty in indicator values which is caused by data inaccuracies and by imperfect validity of bibliometric variables as indicators for theoretical constructs which cannot be directly measured. They can provide transparency about uncertainty. Such a contribution should not be underestimated.

6.2.5 Methodological research in bibliometrics

At present, some part of bibliometric methodology research seems to be caught up in minor, incremental development, improvement, and rediscovery of specific indicators or indicator variants. The bulk of these studies seems to have been written by authors in a rush to propose their own indicator specification without much consideration for the basic reliability and validity of the novel indicator (Bornmann & Marx, 2018). Some topics to which this characterization in my opinion applies are h-index research (Wildgaard, 2019), bibliometric counting methods (Donner, 2020; Gauffriau, 2021), and bibliometric interdisciplinarity measurement (Q. Wang & Schneider, 2020). There seems to be a certain level of dysfunction in the self-organization of this research. Namely, that it is considered a sufficient knowledge contribution to the field when a novel indicator or indicator variant is proposed without so much as to demonstrate improved validity over existing standard indicators (for existing constructs) or basic validity evidence being presented (for newly introduced concepts and indicators). This lenient convention may need to be addressed sooner or later in the interest

of cumulativeness of knowledge. The introduction and adoption of unverified methods may lead to wasteful research – research whose results do not hold up to critical scrutiny because the methods, indicators, tools and so on never actually measured what they alleged to measure (validity) or did so only so poorly as to make the results useless (reliability) (Whitley & Kite, 2012). I recommended that the research community of quantitative science studies adopts the theoretical and methodological apparatus of reliability and validity of psychology and psychometrics (Hunter & Schmidt, 2004) and continues to develop and adapt it for its own needs.

6.2.6 Adoption of explicit uncertainty reporting

A novel framework to calculate uncertainty intervals for arbitrary bibliometric indicators which communicate the uncertainty due to data quality issues was introduced in Chapter 5. For bibliometric studies of critical importance, a further developed model of the introduced novel uncertainty propagation and quantification framework should be applied. This should be built on a solid foundation of empirical error studies for the relevant corpus and data source, which is likely to mean that new data on error distributions has to be collected, covering more metadata fields than investigated in the studies of this project. Moreover, to rule out misguided conclusions from earlier studies, some of them may need to be replicated with the application of a fully worked out bibliometric uncertainty framework that includes data quality and indicator validity as sources of uncertainty.

Table 6.1 gives a concise overview of these and further recommendations organized by the different target groups.

Table 6.1: Summary of recommendations

Indicator	Recommendations
provider/user	
group	

bibliometric data
and analysis
platform providers

- intensify efforts to increase data quality (citation matching, institution affiliation disambiguation, document types, publication months)
- enhance transparency of data quality assurance processes and their results, for example by publishing internal data quality audit reports and benchmark data sets
- offer easily accessible tools for correction and curation of data by user community
- eventually build uncertainty ranges into systems for basic indicators (e.g., ranges of probable citation counts for individual items)
- use month-based citation windows for analysis platforms

bibliometric
research and
evaluation centers

- use month-based citation windows
- continue to conduct independent empirical studies and audits of source data quality, focussing increasingly on open data
- begin to include data quality-based uncertainty ranges for key indicator values
- further develop the framework for uncertainty propagation and expression

research managers,
evaluators, and
other
decision-makers

- build awareness of bibliometric indicator uncertainty and do not base decisions on data and indicator values whose uncertainty from data inaccuracies is unknown
- exercise caution in the interpretation of results if numbers of observations are low and indicator differences are not large
- request uncertainty estimates for key indicators values from data providers and reporting contractors
- report data errors and engage with bibliometric data providers to make corrections

bibliometric
researchers

- exercise caution in the interpretation of results if numbers of observations are low and indicator value differences are not large
- use month-based citation windows
- elevate the standards for the reporting of reliability and validity of bibliometric indicators in research publications
- collect and publish reliability and validity evidence alongside novel indicators and indicator variants
- report data errors and engage with bibliometric data providers to make corrections

6.3 Limitations

The specific limitations of the studies have been discussed separately in the preceding chapters. But there are some broader limitations that apply across

studies that should also be mentioned. The studies presented in the substantive chapters have looked at the practically relevant effects of data quality on indicator values primarily at the intermediate size level of research performing units, namely research groups and organizations. Bibliometric research is also done on the level of individual researchers and on the level of countries or groups of countries, which were not considered in this dissertation but would be relevant extensions for similar research.

Metrics-based assessment of individual researchers, while fiercely contested on methodological grounds, is nevertheless not uncommon. At this level, one of the main advantages of bibliometrics is absent: the statistical robustness of indicator values that large numbers of observations bring about. As individual-level indicator values are inevitably volatile, the influence of inadequate data quality can be expected to be quite large, in other words, their measurement reliability may be quite low. Apart from evaluation purposes, this possibly low reliability may have important ramifications for basic research that are not always recognized. If indicators such as numbers of publications and numbers of citations are affected by measurement errors and are used as proxy variables for unobservable traits of researchers, such as productivity, impact, esteem, or influence, then any associations, like correlations or regression coefficients, will be underestimated because of the random component in the measured values. On the other hand, the much larger number of observations usually available for bibliometric research at the country level would lead one to expect that imperfect data quality may have smaller or even negligible effects here. Yet, this needs to be positively established by empirical research first.

In this dissertation I have argued and demonstrated that poor data significantly impair the reliability and validity of bibliometric indicators. Yet all of this research has been done within the confines of the data that are actually indexed and available in proprietary bibliographic-bibliometric databases. Their coverage of primary research publications is known to be selective and incomplete and the selectivity criteria and their application have been subject to debate (Fenton et al., 2024; Hicks & Wang, 2011; T. N. Van Leeuwen et al., 2001). Nor is the coverage in any way based on random sampling which makes inferential statistics inapplicable as a possible method to generalize from the indexed content to the not indexed content outside of the databases. This confinement to indexed literature masks a major threat to validity that cannot possibly be accounted for in the framework of this research, namely the completeness of the literature in

the available data that is relevant to a given bibliometric study. If the target of a study is as ambitious as all the global literature of basic scientific or scholarly solidity then whatever results one obtains from the major proprietary databases will be highly biased by scientific discipline, by geographical origin of scholarship and by language. Possibly also by the type of publication. For instance, Scopus covers many non-English language journals and many scholarly books but practically all books that it covers are English only³⁷. Thus, it should be noted as a considerable limitation of this research and of other research that similarly relies on these databases, that the estimates of accuracy of data and reliability of indicators are conditioned by this delimited subset of research literature and are necessarily distorted with respect to a representation of the entire global research literature. Such research into database data quality and indicator reliability is always database-specific and cannot simply be generalized beyond the data sources.

The data that are indexed in bibliometric databases are not static, in fact, they are frequently updated. Most updates are newly indexed content but occasionally already indexed metadata are amended or corrected. One would expect that updates to older data are improvements that tend to fix inaccuracies rather than introducing new ones. But this is by no means guaranteed. It is therefore a further limitation that in this dissertation research, the data has been treated from a slightly unrealistic perspective of assuming some longer term stability. It is presently not known if this assumption is justified or not.

Throughout the studies, care was taken to design them in such a way that findings generalize as much as possible, yet some of the results should be considered to generalize better than others. The studies on document types and citation count errors used random samples within specified constraints. Document type inaccuracies and institution affiliation disambiguation were studied across two major commercial bibliometric data sources. Beyond these data sources, error rates for these metadata could be quite different. An important limitation of the study of the affiliation disambiguation systems was that only gold standard data on German institutions was used. The science systems of countries are known to differ in their complexity which likely affects error rates in disambiguation across countries. On the other hand, the results on publication month effects are conceptually independent of data source, scientific discipline, country, and other factors and can be assumed to be reproduce under any conditions.

³⁷See Scopus Book title list available at <https://www.elsevier.com/products/scopus/content>

6.4 Further research directions

From the above-mentioned limitation of my research that it does not fully do justice to the dynamic nature of database content one obvious methodological recommendation for similar studies is to design such studies in a way that makes it easy to repeat them after some time and compare the earlier with the newer results.

There is a certain flexibility in bibliometric practice about which document types to include and exclude in bibliometric studies and in research evaluation and how to handle the interaction of document type and citation count normalization. Some studies and research groups routinely analyze only research articles as assigned in the data source and calculate reference citation count values for normalization on this basis. Others use both ostensible articles and reviews as one undifferentiated kind of substantive research papers and calculate a common normalization score. Yet others may include articles, reviews, and potentially other types such as conference papers as separate categories and treat them distinctly with their own reference scores. For data sources with incomplete or very unreliable document type data, another possibility is the selection of substantive comparable papers based on combinations of criteria such as presence of named authors, a certain page length threshold and number of cited references. It would be an important contribution to systematically study if these different options have a substantial influence on obtained results and if some options are affected more than others by document type inaccuracies.

I have pointed out above the implication of the skewed distribution of missing citations that most publications will have no or few missing citations but a small number of papers will have many missing citations. Table 5.2 showed that of the sample of 372 papers, the majority, 263 papers (71 %), had no missing citations, a further 67 (18 %) had just one missing citation. However, one paper had 15 missing citations and another one had 26. In the first case, this works out to 51 % missing citations and in the second case 65 % missing citations. In my opinion, rather than seeking comfort in the fact that most papers are not affected by substantial numbers of missing citations, the result that missing citations are highly concentrated on a very small fraction of papers should motivate more research into the consequences of this finding. This result is most likely caused by false negatives in citation matching algorithms. One possibility to progress on this is to exploit the results of different citation matching algorithms on

the same data (Franceschini et al., 2013; Olensky et al., 2016; Rodrigues et al., 2025). If these comparisons can identify a sufficiently large number of records with many missing citations, a qualitative analysis of the reasons for matching failure in these cases might produce very valuable insights on the basis of which matching algorithms could be improved.

From the analysis of the original random sample of missing citations in Chapter 5 and various earlier studies using non-representative samples reviewed in that chapter it is clear that citation counts are subject to non-trivial error rates. One problem in developing improved reference matching algorithms, which are the origin of citation counts, is the inevitable trade-off between precision and recall. If an algorithm is made more lenient in order to match more references, such as those with several minor errors, that also means that it will produce more false positives and produce too high citation counts for some papers, as superficially similar reference strings are linked to papers which actually refer to different papers. There seems to me to be one reasonable way to handle this trade-off transparently that citation database providers might want to experiment with. Reference matching algorithms can be designed in a way such that all match candidates that exceed a specific similarity threshold are accepted as matches, where the similarity may be a complex heuristic or a scalar parameter value. Rather than setting one specific threshold (or decision rule) which provides an acceptable balance of precision and recall one could also set two thresholds, one with high recall and still acceptable precision and one with high precision and still acceptable recall (or more than one decision rule). One would then obtain for some papers two different but plausible citation counts, effectively giving a range of plausible citation counts for each publication record which communicates some of the uncertainty inherent in a reference matching algorithm. Since base scholarly data and derived indicator values are uncertain, a reasonable way to move forward is, for every party involved, data producer and users, to always be upfront about which uncertainties they are aware of and how large they are. If a scholarly database uses a reference matching algorithm that cannot determine with complete certainty if one cited reference refers to an indexed document, then this should be visible to users. Citation counts can be displayed as ranges of likely correct values based on fuzzy reference matching results.

The uncertainty propagation framework borrowed from the physical sciences introduced in Chapter 5 can be extended with some further work in an important

direction. While in this dissertation only the type of uncertainty which is attributable to reliability was considered, the uncertainty propagation framework is not restricted to this. Suitably extended, the framework can include data on construct validity (Carmines & Woods, 2005; McDonald, 2005). To this end, the type of outcome quantity has to be changed from directly measurable variables to theoretical latent constructs proxied by one or more measurable variables. The combined assessment of these two cardinal sources of uncertainty, data quality and indicator validity, is proposed as a critically important topic of follow-up research.

The research reported in this dissertation can hopefully contribute to a rational discussion on the opportunities and limitations of bibliometric research methods and the application of bibliometrics to research evaluation. Bibliometric indicators can serve as valid proxies for theoretically and practically relevant concepts if they have been demonstrated to show sufficient construct validity and reliability. I have shown here how much classic bibliometric indicators' reliability depends on data quality and how to propagate and express this reliability component of total indicator uncertainty as uncertainty ranges for indicator values.

References

- Abalkina, A. (2024). Challenges posed by hijacked journals in Scopus. *Journal of the Association for Information Science and Technology*, 75(4), 395–422. <https://doi.org/10.1002/asi.24855>
- Abramo, G. (2024). The forced battle between peer-review and scientometric research assessment: Why the CoARA initiative is unsound. *Research Evaluation*, rvae021. <https://doi.org/10.1093/reseval/rvae021>
- Abramo, G., Cicero, T., & D’Angelo, C. A. (2012). A sensitivity analysis of researchers’ productivity rankings to the time of citation observation. *Journal of Informetrics*, 6(2), 192–201. <https://doi.org/10.1016/j.joi.2011.12.003>
- Abt, H. A. (1992). What fraction of literature references are incorrect? *Publications of the Astronomical Society of the Pacific*, 104(673), 235. <https://doi.org/10.1086/132982>
- Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*, 9(1), 2158244019829575. <https://doi.org/10.1177/2158244019829575>
- Aksnes, D. W., Piro, F. N., & Fossum, L. W. (2023). Citation metrics covary with researchers’ assessments of the quality of their works. *Quantitative Science Studies*, 4(1), 105–126. https://doi.org/10.1162/qss_a_00241
- Alperin, J. P., Muñoz Nieves, C., Schimanski, L. A., Fischman, G. E., Niles, M. T., & McKiernan, E. C. (2019). How significant are the public dimensions of faculty work in review, promotion and tenure documents? *ELife*, 8, e42254. <https://doi.org/10.7554/eLife.42254>
- Álvarez-Bornstein, B., Morillo, F., & Bordons, M. (2017). Funding acknowledgments in the Web of Science: Completeness and accuracy of collected data. *Scientometrics*, 112(3), 1793–1812. <https://doi.org/10.1007/s11192-017-2453-4>
- Alwin, D. F. (2005). Reliability. In K. Kempf-Leonard (Ed.), *Encyclopaedia of social measurement* (Vol. 3, pp. 351–359). Elsevier.
- Aman, V., & van den Besselaar, P. (2024). Authorship regulations in performance-based funding systems and publication behaviour – A case study of German medical faculties. *Journal of Informetrics*, 18(2), 101500. <https://doi.org/10.1016/j.joi.2024.101500>
- Anonymous. (2002). Errors in citation statistics. *Nature*, 415, 101.
- ARC. (2015). *Excellence in Research for Australia (ERA) 2015 Evaluation*

- Handbook*. [http://archive.erc.gov.au/archive/files/ERA/2015/Key%20Documents/ERA](http://archive.erc.gov.au/archive/files/ERA/2015/Key%20Documents/ERA%2015%20Evaluation%20Handbook.pdf) 2015 Evaluation Handbook.pdf
- Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, 1(1), 377–386. https://doi.org/10.1162/qss_a_00019
- Baccini, A., De Nicolao, G., & Petrovich, E. (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLoS One*, 14(9), e0221212. <https://doi.org/10.1371/journal.pone.0221212>
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. ACM Press.
- Barrios, M., Guilera, G., & Gómez-Benito, J. (2013). Impact and structural features of meta-analytical studies, standard articles and reviews in psychology: Similarities and differences. *Journal of Informetrics*, 7(2), 478–486. <https://doi.org/10.1016/j.joi.2013.01.012>
- Batini, C., & Scannapieco, M. (2006). *Data quality. Concepts, methodologies and techniques*. Berlin, Heidelberg: Springer-Verlag. <https://doi.org/10.1007/3-540-33173-5>
- Berlemann, M., & Haucap, J. (2015). Which factors drive the decision to opt out of individual research rankings? An empirical study of academic resistance to change. *Research Policy*, 44(5), 1108–1115. <https://doi.org/10.1016/j.respol.2014.12.002>
- Besançon, L., Cabanac, G., Labbé, C., & Magazinov, A. (2024). Sneaked references: Fabricated reference metadata distort citation counts. *Journal of the Association for Information Science and Technology*. <https://doi.org/10.1002/asi.24896>
- Bhandari, M., Busse, J., Devereaux, P., Montori, V. M., Swiontkowski, M., Tornetta III, P., Einhorn, T. A., Khera, V., & Schemitsch, E. H. (2007). Factors associated with citation rates in the orthopedic literature. *Canadian Journal of Surgery*, 50(2), 119.
- Bimber, B., & Guston, D. H. (1995). Politics by the same means: Government and science in the united states. In *Handbook of science and technology studies, revised edition* (pp. 554–571). SAGE Publications, Inc.
- Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1), 363–376. https://doi.org/10.1162/qss_a_00018
- Bornmann, L., & Marx, W. (2018). Critical rationalism and the search for stan-

- dard (field-normalized) indicators in bibliometrics. *Journal of Informetrics*, 12(3), 598–604. <https://doi.org/10.1016/j.joi.2018.05.002>
- Braun, T., Glänzel, W., & Schubert, A. (1989a). Assessing assessments of British science. Some facts and figures to accept or decline. *Scientometrics*, 15, 165–170.
- Braun, T., Glänzel, W., & Schubert, A. (1989b). Some data on the distribution of journal publication types in the Science Citation Index database. *Scientometrics*, 15, 325–330. <https://doi.org/10.1007/BF02017057>
- Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *College & Research Libraries*, 67(4), 292–303. <https://doi.org/10.5860/crl.67.4.292>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian generalized linear mixed models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Carmines, E. G., & Woods, J. A. (2005). Validity assessment. In K. Kempf-Leonard (Ed.), *Encyclopaedia of social measurement* (Vol. 3, pp. 933–937). Elsevier.
- Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. *Proceedings of the 19th International Conference on Science and Technology Indicators*, 79–86.
- Céspedes, L., Kozłowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimi, S., Ayeni, P., et al. (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*. <https://doi.org/10.1002/asi.24979>
- Chaiworapongsa, T., Espinoza, J., Gotsch, F., Kim, Y. M., Kim, G. J., Goncalves, L. F., Edwin, S., Kusanovic, J. P., Erez, O., Than, N. G., et al. (2008). The maternal plasma soluble vascular endothelial growth factor receptor-1 concentration is elevated in SGA and the magnitude of the increase relates to Doppler abnormalities in the maternal and fetal circulation. *The Journal of Maternal-Fetal & Neonatal Medicine*, 21(1), 25–40. <https://doi.org/10.1080/14767050701832833>
- Chinchilla-Rodríguez, Z., Costas, R., Robinson-García, N., & Larivière, V. (2024). Examining the quality of the corresponding authorship field in Web of Science and Scopus. *Quantitative Science Studies*, 5(1), 76–97. https://doi.org/10.1162/qss_a_00288
- Cioffi, A., Coppini, S., Massari, A., Moretti, A., Peroni, S., Santini, C., &

- Shahidzadeh Asadi, N. (2022). Identifying and correcting invalid citations due to DOI errors in Crossref data. *Scientometrics*, 127(6), 3593–3612. <https://doi.org/10.1007/s11192-022-04367-w>
- Colliander, C., & Ahlgren, P. (2019). Comparison of publication-level approaches to ex-post citation normalization. *Scientometrics*, 120(1), 283–300. <https://doi.org/10.1007/s11192-019-03121-z>
- Corry, L. (2015). *A brief history of numbers*. Oxford: Oxford University Press.
- Cruz-Castro, L., & Sanz-Menendez, L. (2021). What should be rewarded? Gender and evaluation criteria for tenure and promotion. *Journal of Informetrics*, 15(3), 101196. <https://doi.org/10.1016/j.joi.2021.101196>
- Csiszar, A. (2017). The catalogue that made metrics, and changed science. *Nature*, 551(7679), 163–165. <https://doi.org/10.1038/551163a>
- Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*, 130(4), 2475–2492. <https://doi.org/10.1007/s11192-025-05293-3>
- Cuxac, P., Lamirel, J.-C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, 97, 47–58. <https://doi.org/10.1007/s11192-013-1025-5>
- D’Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology*, 62(2), 257–269. <https://doi.org/10.1002/asi.21460>
- D’Angelo, C. A., & van Eck, N. J. (2020). Collecting large-scale publication data at the level of individual researchers: A practical proposal for author name disambiguation. *Scientometrics*, 123, 883–907. <https://doi.org/10.1007/s11192-020-03410-y>
- Daraio, C., Di Leo, S., & Scannapieco, M. (2022). Accounting for quality in data integration systems: A completeness-aware integration approach. *Scientometrics*, 127(3), 1465–1490. <https://doi.org/10.1007/s11192-022-04266-0>
- De Bellis, N. (2009). *Bibliometrics and citation analysis: from the Science Citation Index to cybermetrics*. Lanham, Toronto, Plymouth: Scarecrow Press.
- De Bruin, R., & Moed, H. (1990). The unification of addresses in scientific publications. In L. Egghe & R. Rousseau (Eds.), *Informetrics 1989/90* (pp. 65–78). Amsterdam: Elsevier Science Publishers.
- Dederke, J., Koch, M., & Willemin, S. (2024). The representation of Swiss

- higher education institutions in five bibliometric databases. *Qualität in Der Wissenschaft*, 18(4), 117–124.
- Demetrescu, C., Ribichini, A., & Schaerf, M. (2018). Accuracy of author names in bibliographic data sources: An Italian case study. *Scientometrics*, 117(3), 1777–1791. <https://doi.org/10.1007/s11192-018-2945-x>
- Dickersin, K., Scherer, R., Suci, E. S. T., & Gil-Montero, M. (2002). Problems with indexing and citation of articles with group authorship. *JAMA*, 287(21), 2772–2774. <https://doi.org/10.1001/jama.287.21.2772>
- Dieks, D., & Chang, H. (1976). Differences in impact of scientific publications: Some indices derived from a citation analysis. *Social Studies of Science*, 6(2), 247–267. <https://doi.org/10.1177/030631277600600204>
- Donner, P. (2016a). Missing citations due to exact reference matching: Analysis of a random sample from WoS. Are publications from peripheral countries disadvantaged? *21st International Conference on Science and Technology Indicators-STI 2016. Book of Proceedings*, 85–89.
- Donner, P. (2016b). Enhanced self-citation detection by fuzzy author name matching and complementary error estimates. *Journal of the Association for Information Science and Technology*, 67(3), 662–670. <https://doi.org/10.1002/asi.23399>
- Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, 113(1), 219–236. <https://doi.org/10.1007/s11192-017-2483-y>
- Donner, P. (2020). A validation of coauthorship credit models with empirical data from the contributions of PhD candidates. *Quantitative Science Studies*, 1(2), 551–564. https://doi.org/10.1162/qss_a_00048
- Donner, P., & Blümel, C. (2024). Towards a valid bibliometric measure of epistemic breadth of researchers. *arXiv Preprint arXiv:2411.02005*. <https://doi.org/10.48550/arXiv.2411.02005>
- Donner, P., Rimmert, C., & van Eck, N. J. (2020). Comparing institutional-level bibliometric research performance indicator values based on different affiliation disambiguation systems. *Quantitative Science Studies*, 1(1), 150–170. https://doi.org/10.1162/qss_a_00013
- Dubin, D. (2004). The most influential paper Gerard Salton never wrote. *Library Trends*, 52(4), 748–764.
- Dyachenko, E., & Pislyakov, V. (2010). Scientometric journals in scientometric databases. Microlevel analysis of lost citations. *11th International Conference on Science and Technology Indicators. Book of Abstracts*, 84–86.

- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Dordrecht: Springer Science+Business Media.
- El-Ouahi, J. (2023). The Arabic Citation Index: Toward a better understanding of Arab scientific literature. *Quantitative Science Studies*, 4(3), 728–755. https://doi.org/10.1162/qss_a_00261
- Elzinga, A. (2012). Features of the current science policy regime: Viewed in historical perspective. *Science and Public Policy*, 39(4), 416–428. <https://doi.org/10.1093/scipol/scs046>
- Erman, N., & Todorovski, L. (2015). The effects of measurement error in case of scientific network analysis. *Scientometrics*, 104(2), 453–473. <https://doi.org/10.1007/s11192-015-1615-5>
- European Commission and Directorate-General for Research and Innovation. (2024). *Science, research and innovation performance of the EU, 2024 – A competitive Europe for a sustainable future*. Publications Office of the European Union. <https://doi.org/doi/10.2777/965670>
- Fenton, A., Donner, P., Ambrasat, J., Fabian, G., & Heger, C. (2024). How accurate are Scopus publication counts of researchers? A survey-bibliometric comparison for Germany. *28th International Conference on Science, Technology and Innovation Indicators (STI2024), Berlin, Germany*. <https://doi.org/10.5281/zenodo.13991905>
- Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2013). A novel approach for estimating the omitted-citation rate of bibliometric databases with an application to the field of bibliometrics. *Journal of the American Society for Information Science and Technology*, 64(10), 2149–2156. <https://doi.org/10.1002/asi.22898>
- Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2014). Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? *Journal of Informetrics*, 8(3), 751–765. <https://doi.org/10.1016/j.joi.2014.07.003>
- Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015a). Errors in DOI indexing by bibliometric databases. *Scientometrics*, 102(3), 2181–2186. <https://doi.org/10.1007/s11192-014-1503-4>
- Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015b). Influence of omitted citations on the bibliometric statistics of the major manufacturing journals. *Scientometrics*, 103(3), 1083–1122. <https://doi.org/10.1007/s11192-015-1583-9>
- Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). The museum of

- errors/horrors in Scopus. *Journal of Informetrics*, 10(1), 174–182. <https://doi.org/10.1016/j.joi.2015.11.006>
- French, J. C., Powell, A. L., & Schulman, E. (2000). Using clustering strategies for creating authority files. *Journal of the American Society for Information Science*, 51(8), 774–786. [https://doi.org/10.1002/\(SICI\)1097-4571\(2000\)51:8%3C774::AID-ASI90%3E3.0.CO;2-P](https://doi.org/10.1002/(SICI)1097-4571(2000)51:8%3C774::AID-ASI90%3E3.0.CO;2-P)
- Gadd, E., Holmes, R., & Shearer, J. (2021). Developing a method for evaluating global university rankings. *Scholarly Assessment Reports*, 3(1), 2. <https://doi.org/10.29024/sar.31>
- Gai, S.-S., Liu, X.-L., Zhang, S.-L., & Liu, R.-Y. (2015). The citation evolution law of papers published in the same year but different month. *Learned Publishing*, 28(2), 132–139. <https://doi.org/10.1087/20150207>
- Galvez, C., & Moya-Anegón, F. (2007). Standardizing formats of corporate source data. *Scientometrics*, 70(1), 3–26. <https://doi.org/10.1007/s11192-007-0101-0>
- Gálvez, C., & Moya-Anegón, F. (2006). The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). *Scientometrics*, 69(2), 323–345. <https://doi.org/10.1007/s11192-006-0156-3>
- García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in psychology. *Journal of the American Society for Information Science and Technology*, 61(10), 2070–2085. <https://doi.org/10.1002/asi.21372>
- García-Pérez, M. A. (2011). Strange attractors in the Web of Science database. *Journal of Informetrics*, 5(1), 214–218. <https://doi.org/10.1016/j.joi.2010.07.006>
- García-Pérez, M. A. (2015). Online supplemental information: A sizeable black hole for citations. *Scientometrics*, 102(2), 1655–1659. <https://doi.org/10.1007/s11192-014-1348-x>
- Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159), 108–111. <https://doi.org/10.1126/science.122.3159.108>
- Garfield, E. (1964). Science citation index — a new dimension in indexing. *Science*, 144(3619), 649–654. <https://doi.org/10.1126/science.144.3619.649>
- Garfield, E. (1978). False publication dates and other rip-offs. *Current Contents*, 20, 5–8.
- Garfield, E. (1979). *Citation indexing—its theory and application in science*,

- technology, and humanities. ISI Press, Philadelphia.
- Garfield, E. (1983). Quality control at ISI: A piece of your mind can help us in our quest for error-free bibliographic information. *Current Contents*, 19, 5–12.
- Garfield, E. (1996). When to cite. *The Library Quarterly*, 66(4), 449–458.
- Garfield, E. (2001). Recollections of Irving H. Sher 1924–1996: Polymath/information scientist extraordinaire. *Journal of the American Society for Information Science and Technology*, 52(14), 1197–1202. <https://doi.org/10.1002/asi.1187>
- Gauch, S. (2021). The ironic becomings of reflexivity—The case of citation theory in bibliometrics. *Historical Social Research/Historische Sozialforschung*, 46(2), 155–177. <https://doi.org/10.12759/hsr.46.2021.2.155-177>
- Gauffriau, M. (2021). Counting methods introduced into the bibliometric research literature 1970–2018: A review. *Quantitative Science Studies*, 2(3), 932–975. https://doi.org/10.1162/qss_a_00141
- Giri, R., & Das, A. K. (2011). Indian Citation Index: a new web platform for measuring performance of Indian research periodicals. *Library Hi Tech News*, 28(3), 33–35. <https://doi.org/10.1108/07419051111145154>
- Glänzel, W. (1996). The need for standards in bibliometric research and technology. *Scientometrics*, 35, 167–176. <https://doi.org/10.1007/BF02018475>
- 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. <https://doi.org/10.1080/09737766.2008.10700836>
- Glänzel, W., & Thijs, B. (2004). Does co-authorship inflate the share of self-citations? *Scientometrics*, 61(3), 395–404. <https://doi.org/10.1023/B:SCIE.0000045117.13348.b1>
- Gläser, J. (2024). How can we make “research quality” a theoretical concept? *Research Evaluation*, 33, rva038. <https://doi.org/10.1093/reseval/rvae038>
- Godin, B. (2005). *Measurement and statistics on science and technology: 1920 to the present*. Routledge.
- Gorin, S. V., Koroleva, A. M., Gerasimov, A. N., & Voronov, A. A. (2020). The Russian Science Citation Index (RSCI): the first three years (2016–2018). *European Science Editing*, 46, e51051. <https://doi.org/10.3897/ese.2020.e51051>
- Gorraiz, J., & Schloegl, C. (2008). A bibliometric analysis of pharmacology and pharmacy journals: Scopus versus Web of Science. *Journal of Information Science*, 34(5), 715–725. <https://doi.org/10.1177/0165551507086991>

- Grassano, N., Rotolo, D., Hutton, J., Lang, F., & Hopkins, M. M. (2017). Funding data from publication acknowledgments: Coverage, uses, and limitations. *Journal of the Association for Information Science and Technology*, 68(4), 999–1017. <https://doi.org/10.1002/asi.23737>
- Gunashekar, S., Wooding, S., & Guthrie, S. (2017). How do NIHR peer review panels use bibliometric information to support their decisions? *Scientometrics*, 112, 1813–1835. <https://doi.org/10.1007/s11192-017-2417-8>
- Haddow, G., & Hammarfelt, B. (2019). Quality, impact, and quantification: Indicators and metrics use by social scientists. *Journal of the Association for Information Science and Technology*, 70(1), 16–26. <https://doi.org/10.1002/asi.24097>
- Hager, S., Schwarz, C., & Waldinger, F. (2024). Measuring science: Performance metrics and the allocation of talent. *American Economic Review*, 114(12), 4052–4090. <https://doi.org/10.1257/aer.20230515>
- Hammarfelt, B., Nelhans, G., Eklund, P., & Åström, F. (2016). The heterogeneous landscape of bibliometric indicators: Evaluating models for allocating resources at Swedish universities. *Research Evaluation*, 25(3), 292–305. <https://doi.org/10.1093/reseval/rvv040>
- Hammarfelt, B., & Rushforth, A. D. (2017). Indicators as judgment devices: An empirical study of citizen bibliometrics in research evaluation. *Research Evaluation*, 26(3), 169–180. <https://doi.org/10.1093/reseval/rvx018>
- Hargens, L. L., & Schuman, H. (1990). Citation counts and social comparisons: Scientists' use and evaluation of citation index data. *Social Science Research*, 19(3), 205–221. [https://doi.org/10.1016/0049-089X\(90\)90006-5](https://doi.org/10.1016/0049-089X(90)90006-5)
- Harzing, A.-W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the social sciences? *Scientometrics*, 94(1), 23–34. <https://doi.org/10.1007/s11192-012-0738-1>
- Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics*, 76(1), 169–185. <https://doi.org/10.1007/s11192-007-1892-8>
- Haunschild, R., Daniels, A. D., & Bornmann, L. (2022). Scores of a specific field-normalized indicator calculated with different approaches of field-categorization: Are the scores different or similar? *Journal of Informetrics*, 16(1), 101241. <https://doi.org/10.1016/j.joi.2021.101241>
- Haunschild, R., Schier, H., Marx, W., & Bornmann, L. (2018). Algorithmically generated subject categories based on citation relations: An empirical micro

- study using papers on overall water splitting. *Journal of Informetrics*, 12(2), 436–447. <https://doi.org/10.1016/j.joi.2018.03.004>
- Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2024). Analysis of the publication and document types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar. *arXiv Preprint arXiv:2406.15154*. <https://doi.org/10.48550/arXiv.2406.15154>
- Hauschke, C., & Nazarovets, S. (2025). (Non-) retracted academic papers in OpenAlex. *Journal of Information Science*. <https://doi.org/10.1177/01655515251322478>
- Held, M., Laudel, G., & Gläser, J. (2021). Challenges to the validity of topic reconstruction. *Scientometrics*. <https://doi.org/10.1007/s11192-021-03920-3>
- Heneberg, P. (2013). Supposedly uncited articles of Nobel laureates and Fields medalists can be prevalently attributed to the errors of omission and commission. *Journal of the American Society for Information Science and Technology*, 64(3), 448–454. <https://doi.org/10.1002/asi.22788>
- Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1), 387–395. https://doi.org/10.1162/qss_a_00020
- Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2), 251–261. <https://doi.org/10.1016/j.respol.2011.09.007>
- Hicks, D., & Wang, J. (2011). Coverage and overlap of the new social sciences and humanities journal lists. *Journal of the American Society for Information Science and Technology*, 62(2), 284–294. <https://doi.org/10.1002/asi.21458>
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429–431. <https://doi.org/10.1038/520429a>
- Hua, W. (2001). The development of the Chinese Social Sciences Citation Index. *The Indexer: The International Journal of Indexing*, 22(3), 128–129. <https://doi.org/10.3828/indexer.2001.22.3.6>
- Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*, 1(2), 445–478. https://doi.org/10.1162/qss_a_00031
- Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). Institution name disambiguation for research assessment. *Scientometrics*, 99, 823–838. <https://doi.org/10.1007/s11192-014-1388-8>

[//doi.org/10.1007/s11192-013-1214-2](https://doi.org/10.1007/s11192-013-1214-2)

- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, London, New Delhi: Sage Publications, Inc.
- Irvine, J., Martin, B., Peacock, T., & Turner, R. (1985). Charting the decline in British science. *Nature*, *316*(6029), 587–590.
- Jacsó, P. (2009). Errors of omission and their implications for computing scientometric measures in evaluating the publishing productivity and impact of countries. *Online Information Review*, *33*(2), 376–385. <https://doi.org/10.1108/14684520910951276>
- JCGM. (2008). *Evaluation of measurement data – guide to the expression of uncertainty in measurement* (JCGM 100). Joint Committee for Guides in Metrology.
- Jergas, H., & Baethge, C. (2015). Quotation accuracy in medical journal articles—a systematic review and meta-analysis. *PeerJ*, *3*, e1364. <https://doi.org/10.7717/peerj.1364>
- Jiang, Y., Zheng, H.-T., Wang, X., Lu, B., & Wu, K. (2011). Affiliation disambiguation for constructing semantic digital libraries. *Journal of the American Society for Information Science and Technology*, *62*(6), 1029–1041. <https://doi.org/10.1002/asi.21538>
- Jin, B., & Wang, B. (1999). Chinese Science Citation Database: its construction and application. *Scientometrics*, *45*, 325–332. <https://doi.org/10.1007/BF02458440>
- Jonnalagadda, S., & Topham, P. (2010). NEMO: Extraction and normalization of organization names from PubMed affiliation strings. *Journal of Biomedical Discovery and Collaboration*, *5*, 50.
- Kamrani, P., Dorsch, I., & Stock, W. G. (2021). Do researchers know what the h-index is? And how do they estimate its importance? *Scientometrics*, *126*(7), 5489–5508. <https://doi.org/10.1007/s11192-021-03968-1>
- Kealey, T. (1991). Government-funded academic science is a consumer good, not a producer good: a comparative reassessment of Britain’s scientific and technological achievements since 1794 and a comment on the bibliometry of B. Martin and J. Irvine. *Scientometrics*, *20*(2), 369–394.
- Kealey, T., & Micheals, P. J. (2019). Science and liberty. A complicated relationship. In P. J. Micheals & T. Kealey (Eds.), *Scientocracy. The tangled web of public science and public policy*. Washington, DC: Cato Institute.
- Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates

- the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, 68(4), 984–998. <https://doi.org/10.1002/asi.23734>
- Korn, E. L., & Graubard, B. I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 24, 193–201.
- Krauskopf, E. (2019). Missing documents in Scopus: the case of the journal *Enfermería Nefrológica*. *Scientometrics*, 119(1), 543–547. <https://doi.org/10.1007/s11192-019-03040-z>
- Krauskopf, E., & Salgado, M. (2023). Inconsistency in the registration of the Digital Object Identifier (DOI) of articles on Web of Science and Scopus. *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*, 37(96), 129–144. <https://doi.org/10.22201/iibi.24488321xe.2023.96.58784>
- Kronman, U., Gunnarsson, M., & Karlsson, S. (2010). *The bibliometric database at the Swedish Research Council – Contents, methods and indicators*. Stockholm: Swedish Research Council.
- Leibel, C., & Bornmann, L. (2026). Introducing multiverse analysis to bibliometrics: The case of team size effects on disruptive research. *Quantitative Science Studies*. <https://doi.org/10.1162/QSS.a.475>
- Lenke, C., & Taubert, N. (2024). *Institutional coding*. <https://gitlab.ub.uni-bielefeld.de/ag-bibliometrie/institutional-coding>
- Lenzerini, M., & Daraio, C. (2019). Challenges, approaches and solutions in data integration for research and innovation. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 397–420). Cham: Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-02511-3_15
- Levitt, J. M., & Thelwall, M. (2011). A combined bibliometric indicator to predict article impact. *Information Processing & Management*, 47(2), 300–308. <https://doi.org/10.1016/j.ipm.2010.09.005>
- Lewis, G., Cottrell, R., & Dixon, D. (1999). Bibliometric indicators to assist the peer review process in grant decisions. *Research Evaluation*, 8(1), 47–52. <https://doi.org/10.3152/147154499781777621>
- Leydesdorff, L. (1988). Problems with the “measurement” of national scientific performance. *Science and Public Policy*, 15(3), 149–152.
- Leydesdorff, L. (2001). *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications* (2nd ed.). Universal Publishers.

- Leydesdorff, L. (2017). The positive side of discursive disagreements in the social sciences. *Journal of Informetrics*, 11(4), 1043. <https://doi.org/10.1016/j.joi.2017.09.006>
- Leydesdorff, L., Wouters, P., & Bornmann, L. (2016). Professional and citizen bibliometrics: Complementarities and ambivalences in the development and use of indicators—a state-of-the-art report. *Scientometrics*, 109(3), 2129–2150. <https://doi.org/10.1007/s11192-016-2150-8>
- Liang, L., Zhong, Z., & Rousseau, R. (2014). Scientists’ referencing (mis) behavior revealed by the dissemination network of referencing errors. *Scientometrics*, 101(3), 1973–1986. <https://doi.org/10.1007/s11192-014-1275-x>
- Liu, W. (2020). Accuracy of funding information in scopus: A comparative case study. *Scientometrics*, 124(1), 803–811. <https://doi.org/10.1007/s11192-020-03458-w>
- Liu, W., Hu, G., & Tang, L. (2018). Missing author address information in Web of Science—An explorative study. *Journal of Informetrics*, 12(3), 985–997. <https://doi.org/10.1016/j.joi.2018.07.008>
- Liu, W., Huang, M., & Wang, H. (2021). Same journal but different numbers of published records indexed in Scopus and Web of Science Core Collection: Causes, consequences, and solutions. *Scientometrics*, 126(5), 4541–4550. <https://doi.org/10.1007/s11192-021-03934-x>
- Locke, E. A. (2012). Construct validity vs. Concept validity. *Human Resource Management Review*, 22(2), 146–148. <https://doi.org/10.1016/j.hrmr.2011.11.008>
- Lohr, S. L. (2010). *Sampling: Design and analysis*. Brooks/Cole, Cengage Learning.
- Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 1–19.
- Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics*, 1(2), 145–154. <https://doi.org/10.1016/j.joi.2006.09.007>
- Maisano, D., Mastrogiacomo, L., Ferrara, L., & Franceschini, F. (2025). A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. *Scientometrics*, 1–38. <https://doi.org/10.1007/s11192-025-05244-y>
- Martin, B. (1991). The bibliometric assessment of UK scientific performance a reply to Braun, Glänzel and Schubert. *Scientometrics*, 20(2), 333–357. <https://doi.org/10.1007/BF02017524>
- Martin, B. (1994). British science in the 1980s—has the relative decline contin-

- ued? *Scientometrics*, 29(1), 27–56. <https://doi.org/10.1007/BF02018382>
- Martin, B. R. (2017). When social scientists disagree: Comments on the Butler-van den Besselaar debate. *Journal of Informetrics*, 11(3), 937–940. <https://doi.org/10.1016/j.joi.2017.05.021>
- Martin, B. R., Irvine, J., Narin, F., & Sterritt, C. (1987). The continuing decline of British science. *Nature*, 330(6144), 123–126.
- McDonald, M. P. (2005). Validity, data sources. In K. Kempf-Leonard (Ed.), *Encyclopaedia of social measurement* (Vol. 3, pp. 939–948). Elsevier.
- McKiernan, E. C., Schimanski, L. A., Nieves, C. M., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Meta-research: Use of the journal impact factor in academic review, promotion, and tenure evaluations. *ELife*, 8, e47338. <https://doi.org/10.7554/eLife.47338>
- Mehrad, J., & Arastoopoor, S. (2012). Islamic World Science Citation Center (ISC): Evaluating scholarly journals based on citation analysis. *Acta Informatica Medica*, 20(1), 40. <https://doi.org/10.5455/aim.2012.20.40-43>
- Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer. <https://doi.org/10.1007/1-4020-3714-7>
- Moed, H. F., & van Leeuwen, T. N. (1995). Improving the accuracy of Institute for Scientific Information's journal impact factors. *Journal of the American Society for Information Science*, 46(6), 461–467. [https://doi.org/10.1002/\(SICI\)1097-4571\(199507\)46:6%3C461::AID-ASI5%3E3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-4571(199507)46:6%3C461::AID-ASI5%3E3.0.CO;2-G)
- Moed, H. F., & Vriens, M. (1989). Possible inaccuracies occurring in citation analysis. *Journal of Information Science*, 15(2), 95–107. <https://doi.org/10.1177/016555158901500205>
- Moed, H., De Bruin, R., & 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. <https://doi.org/10.1007/BF02017338>
- Moed, H., & van Raan, A. (1988). Indicators of research performance: Applications in university research policy. In *Handbook of quantitative studies of science and technology* (pp. 177–192). Amsterdam: Elsevier Science Publishers.
- Mokhnacheva, Y. V. (2023). Document types indexed in WoS and Scopus: similarities, differences, and their significance in the analysis of publication activity. *Scientific and Technical Information Processing*, 50(1), 40–46. <https://doi.org/10.3103/S0147688223010033>
- Montesi, M., & Mackenzie Owen, J. (2008). Research journal articles as docu-

- ment genres: Exploring their role in knowledge organization. *Journal of Documentation*, 64(1), 143–167. <https://doi.org/10.1108/00220410810844196>
- National Science Board. (1973). *Science Indicators 1972*. National Science Foundation, Washington, DC. <http://files.eric.ed.gov/fulltext/ED084150.pdf>
- Nederhof, A. J. (1988). The validity and reliability of evaluation of scholarly performance. In *Handbook of quantitative studies of science and technology* (pp. 193–228). Amsterdam: Elsevier Science Publishers.
- Negishi, M., Sun, Y., & Shigi, K. (2004). Citation database for Japanese papers: A new bibliometric tool for Japanese academic society. *Scientometrics*, 60(3), 333–351. <https://doi.org/10.1023/B:SCIE.0000034378.38698.b2>
- Okubo, Y. (1997). *Bibliometric indicators and analysis of research systems: Methods and examples* (No. 1997/01; OECD Science, Technology and Industry Working Papers). OECD. <https://doi.org/10.1787/208277770603>
- Olenksy, M. (2015). *Data accuracy in bibliometric data sources and its impact on citation matching* [PhD thesis, Humboldt-Universität zu Berlin, Philosophische Fakultät I]. <https://doi.org/10.18452/17122>
- Olenksy, M., Schmidt, M., & van Eck, N. J. (2016). Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of Science. *Journal of the Association for Information Science and Technology*, 67(10), 2550–2564. <https://doi.org/10.1002/asi.23590>
- Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & Delgado López-Cózar, E. (2017). The lost academic home: institutional affiliation links in Google Scholar Citations. *Online Information Review*, 41(6), 762–781. <https://doi.org/10.1108/OIR-10-2016-0302>
- Patsopoulos, N. A., Analatos, A. A., & Ioannidis, J. P. (2005). Relative citation impact of various study designs in the health sciences. *JAMA*, 293(19), 2362–2366. <https://doi.org/10.1001/jama.293.19.2362>
- Peritz, B. C. (1983). Are methodological papers more cited than theoretical or empirical ones? The case of sociology. *Scientometrics*, 5, 211–218. <https://doi.org/10.1007/BF02019738>
- Peroni, S., & Shotton, D. (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, 1(1), 428–444. https://doi.org/10.1162/qss_a_00023
- Polanyi, M. (1962). The republic of science: Its political and economic theory. *Minerva*, 1(1), 54–73.
- Porter, A. L. (1977). Citation analysis: Queries and caveats. *Social Studies of*

- Science*, 7(2), 257–267. <https://doi.org/10.1177/030631277700700207>
- Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today’s academic world. *Publications*, 9(1), 12. <https://doi.org/10.3390/publications9010012>
- Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators, STI 2022* (p. sti22193). <https://doi.org/10.5281/zenodo.6936227>
- Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, 25(4), 348–349.
- Reyes-Elizondo, A., Calero-Medina, C., Visser, M., & Waltman, L. (2016). *The challenge of identifying universities for a global university ranking [blog post]*. <https://www.cwts.nl/blog?article=n-q2w264>
- Reymert, I. (2021). Bibliometrics in academic recruitment: A screening tool rather than a game changer. *Minerva*, 59(1), 53–78. <https://doi.org/10.1007/s11024-020-09419-0>
- Rice, D. B., Raffoul, H., Ioannidis, J. P., & Moher, D. (2020). Academic criteria for promotion and tenure in biomedical sciences faculties: Cross sectional analysis of international sample of universities. *British Medical Journal*, 369. <https://doi.org/10.1136/bmj.m2081>
- Rimmert, C., Schwechheimer, H., & Winterhager, M. (2017). *Disambiguation of author addresses in bibliometric databases - technical report*. Bielefeld: Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I²SoS). <https://pub.uni-bielefeld.de/record/2914944>
- Robinson-Garcia, N., Torres-Salinas, D., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., & Gorraiz, J. (2023). Errors of measurement in scientometrics: Identification and calculation of systematic errors. *Proceedings of ISSI 2023 – the 19th International Conference of the International Society for Scientometrics and Informetrics*, 2, 387–393. <https://doi.org/10.5281/zenodo.8428899>
- Rodrigues, D., Lopes, A. L., & Batista, F. (2023). Web of Science citation gaps: An automatic approach to detect indexed but missing citations. *12th Symposium on Languages, Applications and Technologies (SLATE 2023)*. <https://doi.org/10.4230/OASlcs.SLATE.2023.5>
- Rodrigues, D., Lopes, A., & Batista, F. (2025). Detecting incoherent citation data among three bibliometric platforms: OpenAlex, Sco-

- pus and Web of Science. *Journal of Information Science*. <https://doi.org/10.1177/01655515251330579>
- Romero, A., Cortés, J., Escudero, C., López, J., & Moreno, J. (2009). Measuring the influence of clinical trials citations on several bibliometric indicators. *Scientometrics*, 80(3), 747–760. <https://doi.org/10.1007/s11192-008-2108-6>
- Rousseau, R., Egghe, L., & Guns, R. (2018). *Becoming metric-wise: A bibliometric guide for researchers*. Cambridge, Kidlington: Chandos Publishing.
- Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, 9(1), 102–117. <https://doi.org/10.1016/j.joi.2014.11.010>
- Rushforth, A., & de Rijcke, S. (2015). Accounting for impact? The Journal Impact Factor and the making of biomedical research in the Netherlands. *Minerva*, 53(2), 117–139. <https://doi.org/10.1007/s11024-015-9274-5>
- Rushforth, A., & Hammarfelt, B. (2023). The rise of responsible metrics as a professional reform movement: A collective action frames account. *Quantitative Science Studies*, 4(4), 879–897. https://doi.org/10.1162/qss_a_00280
- Sargsyan, S., Mirzoyan, A., & Astsatryan, H. (2022). Towards the Armenian Science Citation Index. *PUBMET2022: The 9th Conference on Scholarly Communication in the Context of Open Science*, 80–81.
- Sauvayre, R. (2022). Misreferencing practice of scientists: Inside researchers’ sociological and bibliometric profiles. *Social Epistemology*, 66(6), 719–730. <https://doi.org/10.1080/02691728.2021.2022807>
- Scheidsteger, T., Haunschild, R., & Bornmann, L. (2025). How similar are field-normalized citation impact scores obtained from OpenAlex and three popular commercial databases? An empirical comparison based on large German universities. *Scientometrics*, 1–33. <https://doi.org/10.1007/s11192-025-05338-7>
- Schmidt, F. (2018). *Fehlerabschätzungen bei bibliometrischen Analysen* [Master’s thesis, Technische Hochschule Köln]. <https://publiscologne.th-koeln.de/frontdoor/index/index/docId/1312>
- Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics*, 7(1), 50–62. <https://doi.org/10.1016/j.joi.2012.08.005>
- Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, 9, 281–291. <https://doi.org/10.1007/BF02017249>

- Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, 5(1), 59–73. <https://doi.org/10.1007/BF02097178>
- Schubert, A., Glänzel, W., & Braun, T. (1988). Against absolute methods: Relative scientometric indicators and relational charts as evaluation tools. In A. F. J. van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 137–176). Amsterdam: Elsevier Science Publishers.
- Schulz, J. (2016). Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analyses. *Scientometrics*, 107(3), 1283–1298. <https://doi.org/10.1007/s11192-016-1892-7>
- Selivanova, I. V., Kosyakov, D. V., & Guskov, A. E. (2019). The impact of errors in the Scopus database on the research assessment. *Scientific and Technical Information Processing*, 46, 204–212. <https://doi.org/10.3103/S0147688219030109>
- Sher, I. H., Garfield, E., & Elias, A. W. (1966). Control and Elimination of Errors in ISI Services. *Journal of Chemical Documentation*, 6(3), 132–135. <https://doi.org/10.1021/c160022a0049>
- Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, 13(1), 202–225. <https://doi.org/10.1016/j.joi.2018.12.005>
- Sigogneau, A. (2000). An analysis of document types published in journals related to physics: Proceeding papers recorded in the Science Citation Index database. *Scientometrics*, 47(3), 589–604. <https://doi.org/10.1023/A:1005628218890>
- Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, 354(6312), aaf5239. <https://doi.org/10.1126/science.aaf5239>
- Šipka, P. (2005). The Serbian citation index: Context and content. In P. Ingwersen & B. Larsen (Eds.), *Proceedings of ISSI 2005 - the 10th International Conference of the International Society for Scientometrics and Informetrics* (Vol. 2, pp. 710–711). Karolinska University Press.
- Sirtes, D. (2012). How (dis-) similar are different citation normalizations and the fractional citation indicator? (And how it can be improved). *Proceedings of 17th International Conference on Science and Technology Indicators (STI)*, 894–889.
- Sivertsen, G., & Rushforth, A. (2025). The ongoing reform of research assessment. In G. Sivertsen & L. Langfeldt (Eds.), *Challenges in research*

- policy: Evidence-based policy briefs with recommendations* (pp. 47–54). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-69580-3_7
- Spodick, D. H., & Goldberg, R. J. (1983). The editor’s correspondence: Analysis of patterns appearing in selected specialty and general journals. *The American Journal of Cardiology*, 52(10), 1290–1292. [https://doi.org/10.1016/0002-9149\(83\)90590-8](https://doi.org/10.1016/0002-9149(83)90590-8)
- Stahlschmidt, S., & Schmidt, M. (2019). Measurement variation in bibliometric impact indicators. In G. Catalano, C. Daraio, M. Gregori, H. F. Moed, & G. Ruocco (Eds.), *Proceedings of the 17th Conference of the International Society for Scientometrics and Informetrics (ISSI 2019)* (Vol. 2, pp. 1500–1506). Edizioni Efesto.
- Stahlschmidt, S., & Stephen, D. (2022). From indexation policies through citation networks to normalized citation impacts: Web of Science, Scopus, and Dimensions as varying resonance chambers. *Scientometrics*, 127(5), 2413–2431. <https://doi.org/10.1007/s11192-022-04309-6>
- Struck, A. (2025). *Dependencies in topic delineation and tracking* [PhD thesis, Humboldt-Universität zu Berlin]. <https://doi.org/10.18452/30669>
- Suleimenov, E., Frolova, V., Rog, O., Khasenova, S., Seitkalieva, Z., & Kubieva, T. S. (2009). Creating the Kazakh science citation index. *Scientific and Technical Information Processing*, 36(3), 151–155. <https://doi.org/10.3103/S0147688209030046>
- Swedish Research Council. (2017). *The bibliometric database at the Swedish Research Council – Contents, methods and indicators. Version 2.1*. <https://www.vr.se/download/18.514d156f1639984ae0789d34/1529480565512/The+bibliometric+database+at+the+Swedish+Research+Council+-+contents,+methods+and+indicators.pdf>
- Taylor, J. R. (1997). *An introduction to error analysis* (2nd ed.). Sausalito: University Science Books.
- Teixeira da Silva, J. A. (2024). The inaccurate representation of an author’s publishing name, and impact on reference accuracy. *Scientometrics*, 1–10. <https://doi.org/10.1007/s11192-024-05029-9>
- Thelwall, M. (2016). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics*, 10(2), 336–346. <https://doi.org/10.1016/j.joi.2015.12.007>
- Thelwall, M., Kousha, K., Makita, M., Abdoli, M., Stuart, E., Wil-

- son, P., & Levitt, J. (2023). In which fields do higher impact journals publish higher quality articles? *Scientometrics*, 3915–3933. <https://doi.org/10.1007/s11192-023-04735-0>
- Tierney, E., O’Rourke, C., & Fenton, J. (2015). What is the role of “the letter to the editor”? *European Archives of Oto-Rhino-Laryngology*, 272, 2089–2093. <https://doi.org/10.1007/s00405-014-3289-7>
- Tiesinga, E., Mohr, P. J., Newell, D. B., & Taylor, B. N. (2021). CODATA recommended values of the fundamental physical constants: 2018. *Journal of Physical and Chemical Reference Data*, 50(3), 033105. <https://doi.org/10.1063/5.0064853>
- Tong, S., Chen, F., Yang, L., & Shen, Z. (2023). Novel utilization of a paper-level classification system for the evaluation of journal impact: An update of the CAS Journal Ranking. *Quantitative Science Studies*, 4(4), 960–975. https://doi.org/10.1162/qss_a_00270
- Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1), 5233. <https://doi.org/10.1038/s41598-019-41695-z>
- Tunger, D., Haustein, S., Ruppert, L., Luca, G., & Unterhalt, S. (2010). "The Delphic Oracle" - An Analysis of Potential Error Sources in Bibliographic Databases. *11th International Conference on Science and Technology Indicators. Book of Abstracts*, 282–283.
- Tüür-Fröhlich, T. (2016). *The non-trivial effects of trivial errors in scientific communication and evaluation*. Glückstadt: vwh Verlag Werner Hülsbusch.
- Valderrama-Zurián, J.-C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics*, 9(3), 570–576. <https://doi.org/10.1016/j.joi.2015.05.002>
- Van Eck, N. J., & Waltman, L. (2017). Accuracy of citation data in Web of Science and Scopus. *Proceedings of the 16th International Conference of the International Society for Scientometrics and Informetrics*, 1087–1092. <https://arxiv.org/abs/1906.07011>
- Van Eck, N. J., Waltman, L., van Raan, A. F., Klautz, R. J., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PloS One*, 8(4), e62395. <https://doi.org/10.1371/journal.pone.0062395>
- Van Leeuwen, T. N., Moed, H. F., Tijssen, R. J., Visser, M. S., & van Raan, A. F. (2001). Language biases in the coverage of the science citation index and its

- consequences for international comparisons of national research performance. *Scientometrics*, 51(1), 335–346. <https://doi.org/10.1023/A:1010549719484>
- Van Leeuwen, T., Costas, R., Calero-Medina, C., & Visser, M. (2013). The role of editorial material in bibliometric research performance assessments. *Scientometrics*, 95, 817–828. <https://doi.org/10.1007/s11192-012-0904-5>
- Van Leeuwen, T., van der Wurff, L., & De Craen, A. (2007). Classification of “research letters” in general medical journals and its consequences in bibliometric research evaluation processes. *Research Evaluation*, 16(1), 59–63. <https://doi.org/10.3152/095820207X196786>
- Van Raan, A. (2019). Measuring science: Basic principles and applications of advanced bibliometrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 69–94). Cham: Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-02511-3_10
- Vinkler, P. (2010). *The evaluation of research by scientometric indicators*. Oxford, Cambridge, New Delhi: Chandos Publishing.
- Wade, N. (1975). Citation analysis: A new tool for science administrators. *Science*, 188(4187), 429–432. <https://doi.org/10.1126/science.188.4187.429>
- a
- Wager, E., & Middleton, P. (2008). Technical editing of research reports in biomedical journals. *Cochrane Database of Systematic Reviews*, 4. <https://doi.org/10.1002/14651858.MR000002.pub3>
- Walker, R. L., Sykes, L., Hemmelgarn, B. R., & Quan, H. (2010). Authors’ opinions on publication in relation to annual performance assessment. *BMC Medical Education*, 10, 1–6. <https://doi.org/10.1186/1472-6920-10-21>
- Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2), 365–391. <https://doi.org/10.1016/j.joi.2016.02.007>
- Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C., Tijssen, R. J., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F., Visser, M. S., & 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. <https://doi.org/10.1002/asi.22708>
- Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378–2392. <https://doi.org/10.1002/asi.22748>

- Waltman, L., & van Eck, N. J. (2019). Field normalization of scientometric indicators. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 281–300). Springer.
- Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1), 37–47. <https://doi.org/10.1016/j.joi.2010.08.001>
- Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, 94(3), 851–872. <https://doi.org/10.1007/s11192-012-0775-9>
- Wang, Q., & Schneider, J. W. (2020). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies*, 1(1), 239–263. https://doi.org/10.1162/qss_a_00011
- Waytowich, V. L., Onwuegbuzie, A. J., & Jiao, Q. G. (2006). Characteristics of doctoral students who commit citation errors. *Library Review*. <https://doi.org/10.1108/00242530610655993>
- Whitley, B. E., & Kite, M. E. (2012). *Principles of research in behavioral science* (3rd ed.). New York, London: Routledge.
- Wildgaard, L. (2019). An overview of author-level indicators of research performance. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 361–396). Cham: Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-02511-3_24
- Winterhager, M., Schwechheimer, H., & Rimmert, C. (2014). Institutionenkodierung als Grundlage für bibliometrische Indikatoren. *Bibliometrie - Praxis Und Forschung*, 3. <https://doi.org/10.5283/bpf.209>
- Woelert, P., & McKenzie, L. (2018). Follow the money? How Australian universities replicate national performance-based funding mechanisms. *Research Evaluation*, 27(3), 184–195. <https://doi.org/10.1093/reseval/rvy018>
- Wouters, P. F. (1999). *The citation culture* [PhD thesis, Universiteit van Amsterdam]. <https://dare.uva.nl/search?identifier=b101b769-100f-43e5-b8d2-cac6c11e5bbf>
- Xu, S., Hao, L., An, X., Zhai, D., & Pang, H. (2019). Types of DOI errors of cited references in Web of Science with a cleaning method. *Scientometrics*, 120, 1427–1437. <https://doi.org/10.1007/s11192-019-03162-4>
- Zacharewicz, T., Lepori, B., Reale, E., & Jonkers, K. (2019). Performance-based research funding in EU Member States—a comparative assessment. *Science*

- and Public Policy*, 46(1), 105–115. <https://doi.org/10.1093/scipol/scy041>
- Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*, 129(10), 5869–5891. <https://doi.org/10.1007/s11192-023-04923-y>
- Zhu, J., Hu, G., & Liu, W. (2019). DOI errors and possible solutions for Web of Science. *Scientometrics*, 118, 709–718. <https://doi.org/10.1007/s11192-018-2980-7>
- Zuccala, A., & van Leeuwen, T. (2011). Book reviews in humanities research evaluations. *Journal of the American Society for Information Science and Technology*, 62(10), 1979–1991. <https://doi.org/10.1002/asi.21588>

English language summary

Bibliometric uncertainty: Data quality, reliability, and validity

Bibliometric methods and bibliometric indicators are important, albeit disputed, tools in empirical studies of science systems and in research evaluation. Their validity and acceptance is partly due to their statistical scope, allowing for analyses at large scales, such as entire countries, for which the cost of employing qualitative methods is too high. Bibliometric methods are deeply data-driven which makes the trustworthiness of the statements and interpretations derived from bibliometric studies very data-dependent. Science studies researchers have long recognized that poor quality data is a possible threat to informative scientific outcomes and solid, defensible quantitative assessment of research. They have consequently put considerable effort into empirically investigating the quality of their data although this has not yet precipitated the emergence of a consolidated systematic research program in its own right. Scientometricians and bibliometricians have studied such data quality topics and author name disambiguation, organization name disambiguation, completeness and accuracy of various bibliographic fields such as cited references, identifiers (DOIs), and assigned document types and categories of subject classification systems. Not infrequently, the practical consequences of the observed error rates have remained unclear. How to express the amount of uncertainty in bibliometric that are caused by inaccuracies in the basic data has also remained elusive. The research of this dissertation extends the knowledge on error rates of specific data attributes, namely document types and citation rates, and studies also the performance of

different organization disambiguation systems. It further examines the standard methodology of citation analysis of using nominal year-based citation windows and find this to be biased. In all studies, emphasis is put on the real-world implications of these deficiencies and possible improvements. The final part introduces a novel framework for the propagation and reporting of uncertainties that arise from such data quality issues. I show how inaccuracies in input data can be used to derive realistic error margins for bibliometric indicator result values that quantify their uncertainty due to statistical error in the data. The main result is that bibliometric indicator values are commonly affected by the error sources that were studied that small to moderate differences in indicator values—even for aggregate units such as research groups, departments, universities—might be attributed to data inaccuracies rather than actual performance differences. The implications and consequences of these findings are discussed and recommendations for improvements of current practice are derived.

Nederlandstalige samenvatting

Bibliometrische onzekerheid: gegevenskwaliteit, betrouwbaarheid en validiteit

Bibliometrische methoden en bibliometrische indicatoren zijn belangrijke, zij het omstreden, hulpmiddelen in empirische studies van wetenschapssystemen en in de evaluatie van onderzoek. Hun validiteit en acceptatie zijn deels te danken aan hun statistische reikwijdte, waardoor analyses op grote schaal mogelijk zijn, zoals op nationaal niveau, waarvoor de kosten van het gebruik van kwalitatieve methoden te hoog zouden zijn. Bibliometrische methoden zijn sterk datagedreven. De betrouwbaarheid van de uitspraken en interpretaties die uit bibliometrische studies voortvloeien zijn dan ook sterk afhankelijk van data. Wetenschappers erkennen al lang dat data van slechte kwaliteit een bedreiging vormen voor informatieve wetenschappelijke resultaten en een solide, verdedigbare kwantitatieve beoordeling van onderzoek. Ze hebben daarom aanzienlijke inspanningen geleverd om de kwaliteit van hun data empirisch te onderzoeken. Dit heeft echter nog niet geleid tot de opkomst van een systematisch onderzoeksprogramma. Scientometristen en bibliometristen hebben de datakwaliteit bestudeerd, evenals het doorverwijzen van auteursnamen, organisatienamen, de volledigheid en nauwkeurigheid van verschillende bibliografische velden zoals

geciteerde referenties, identifiers (DOI's) en toegewezen documenttypen en de gehanteerde categorieën van classificatiesystemen van onderwerpen.

Niet zelden zijn de praktische gevolgen van de waargenomen foutpercentages onduidelijk gebleven. Hoe de mate van onzekerheid in bibliometrische gegevens, veroorzaakt door onnauwkeurigheden in de basisgegevens, moet worden uitgedrukt, is eveneens ongrijpbaar gebleven.

Het onderzoek in dit proefschrift breidt de kennis uit over foutpercentages van specifieke dataattributen, namelijk documenttypen en citatiepercentages, en bestudeert ook de prestaties van verschillende systemen voor het ondubbelzinnig maken van de verwijzingen naar organisaties. Het onderzoekt verder de standaardmethodologie voor citatieanalyse met behulp van nominale jaargebaseerde citatievensters en concludeert dat deze bevooroordeeld is. In alle studies wordt de nadruk gelegd op de praktische implicaties van deze tekortkomingen en mogelijke verbeteringen.

Het laatste deel introduceert een nieuw raamwerk voor de verspreiding en rapportage van onzekerheden die voortvloeien uit dergelijke problemen met de datakwaliteit. Het proefschrift laat zien hoe onnauwkeurigheden in invoergegevens kunnen worden gebruikt om realistische foutenmarges af te leiden voor bibliometrische indicatorresultaatwaarden.

Het belangrijkste resultaat is dat bibliometrische indicatorwaarden vaak worden beïnvloed door de onderzochte foutenbronnen, waardoor kleine tot matige verschillen in indicatorwaarden – zelfs voor geaggregeerde eenheden zoals onderzoeksgroepen, afdelingen en universiteiten – mogelijk te wijten zijn aan onnauwkeurigheden in de gegevens in plaats van aan daadwerkelijke prestatieverschillen. De implicaties en gevolgen van deze bevindingen worden besproken en er worden aanbevelingen gedaan voor verbeteringen vande huidige praktijk.

Curriculum Vitae

Paul Donner was born in 1985 in Berlin. Between 1992 and 2004 he attended primary and secondary school in Berlin and Dresden. He studied Library and Information Science, earning a BA (2008) and an MA (2012) from Humboldt-Universität zu Berlin. Donner has been a staff researcher at German Centre for Higher Education Research and Science Studies (DZHW) (up to 2016 called

iFQ). Mr Donner worked in various basic and applied research projects and commercial contract projects for science institutions requiring bibliometric evaluation services and expertise. Mr Donner's main research topics are bibliometric data quality and its consequences for research evaluation, construction of bibliometric measurement methods of theoretically motivated epistemic and social constructs, and validation of bibliometric core methods like citation count normalization, publication credit attribution, and topical publication clustering. Non-research work includes responsibility for data quality assurance and the complete redesign of the relational database for principal common bibliometric data consortium in Germany. Mr Donner also did numerous service projects in the form of provision of custom bibliometric data and research evaluation reports for universities, research institutes, commercial companies, learned societies, advisory bodies, and foundations. He pursued his PhD studies at CWTS, Faculty of Social and Behavioural Sciences, Universteit Leiden.

