



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

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 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.