What is the Required Level of Data Cleaning? A Research Evaluation Case

Peter van den Besselaar[1]* and Ulf Sandström[2]

[1]Department of Organization Sciences and Network Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[2]Department of INDEK, KTH Royal Institute of Technology, Stockholm.

ABSTRACT

Bibliometric methods depend heavily on the quality of data, and cleaning and disambiguating data are very time-consuming. Therefore, quite some effort is devoted to the development of better and faster tools for disambiguating of the data (e.g., Gurney *et al*. 2012). Parallel to this, one may ask to what extent data cleaning is needed, given the intended use of the data. To what extent is there a trade-off between the type of questions asked and the level of cleaning and disambiguating required? When evaluating individuals, a very high level of data cleaning is required, but for other types of research questions, one may accept certain levels of error, as long as these errors do not correlate with the variables under study. In this paper, we present an earlier case study with a rather crude way of data handling as it was expected that the unavoidable error would even out. In this paper, we do a sophisticated data cleaning and disambiguation of the same dataset, and then do the same analysis as before. We compare the results and discuss conclusions about required data cleaning.

Keywords: Coupling data sets, Data cleaning disambiguation, Data error.

INTRODUCTION

Data quality in bibliometric research is often problematic. Synonyms, homonyms, misspellings, and processing errors: this all leads to an error in data. Cleaning and disambiguating are very resource intensive. Quite some research activities are taking place, mainly by information scientists and computer scientists in order to obtain better disambiguation tools and techniques.[1-3] However, the available techniques are still far from perfect (or even optimal). Therefore, it is an important question what kind of data error exists in a specific data set and whether that data error is influencing the outcomes of studies deploying those data. Of course, analytically this is clear: errors that are unrelated to the variables taken into account in the research project are unproblematic. For clarity, our main interest concerns the difference between systematic failures and random failures. The discussion that followed Newman's classical paper (2001)[4] could illustrate this distinction. Newman investigated networks of co-authors in four disciplinary areas and reported some interesting results where he invoked the assumption that the number of unique authors could be identified using two methods (1) first initial disambiguation and (2) all initial disambiguation, in order to establish the lower and upper bound of the numbers of actual authors. This method led Newman to the conclusion that the errors were marginal or of an order of a few percent. As noted by Kim and Diesner[5] this finding has been frequently cited, the paper has received almost 3500 Google Scholar citations, and a

***Address for correspondence:**
E-mail: p.a.a.vanden.besselaar@vu.nl.

Access this article online

Official Publication of

Phcog.Net
Bringing Medicinal Plants Researchers Together

Website: www.jscires.org

DOI: 10.5530/jscires.5.1.3

How to cite this article: Besselaar P, Sandström U. What is the required level of data cleaning? A research evaluation case. J Scientometric Res. 2016;5(1):7-12.

number of scholars have taken for granted that simple disambiguation methods can work quite efficiently. That last proposition has been challenged by Fegley and Torvik[6] for very large datasets and by Kim and Diesner[5] who shows that also in smaller datasets there might be substantial error rates. The latter authors indicate, by the use of more advanced methods for disambiguation that take several types of information into account,[1] that the rate of error can be substantial, especially in areas where there are high numbers of Asian named authors.

While these types of systematic errors are easy to understand as soon as they are demonstrated, but there are others that might be of a different type as they only randomly affect the results. It may be useful to spend more effort into answering these questions as the outcomes may help researchers to decide to what level data cleaning is in fact needed.

THE CASE

An earlier study[7] on the social science division of a research council found that applicants out performnonapplicants and that successful applicants perform better than nonselected applicants (NS) and that the best performing nonsuccessful applicants (BPNS) tend to perform on average at least as good as the successful applicants (S), actually leading to high percentages of false positives and false negatives. Best performing can be defined in different ways. For example in terms of the number of publications, the number of citations or anyother (combination of) indicator(s), such as field normalized indicators of the H-index. Here we use the number of citations for identifying the BPNS applicants.

To come to these conclusions, the applicants' data had to be combined with Web of Science (WoS) data in order to measure the performance of the applicants. Eventually, the data collection had five characteristics that produced the following types of error in the data: (1) the coupling with WoS data on publications and citations was done in a rather crude way: through the family name and first initial, without further cleaning and disambiguating, (2) the analysis was based on SocSCI publications, and publications in science journals by applicants in the sample were not included, (3) only papers with a Dutch address were included and no publications of the applicants that were written when staying abroad without any coauthor residing in the Netherlands, (4) no field normalization was deployed because the authors lacked at that moment data and tools to do so, and (5) other types of scholarly output than WoS in WoS-indexed journals were not included, such as books and chapters, despite the fact that they still play a large role in the social sciences. This was done because the council under study wanted to include only WoS-indexed journals; this against the background that focuses on journals was increasingly characterizing the Dutch system, also the social science fields under study. These were, in fact, the quality indicators the council was interested in and they are increasingly dominant in the science system, also in the social sciences in the Netherlands. As a test, the analysis was done separately for those fields that are strongly journal-dominated (economics, psychology), with similar results as for all social sciences.

All these decisions, of course, result in errors: possible overestimation of performance (name homonyms, fields with high citation and publication levels, and authors focusing only on journal articles) and underestimation (name synonyms, composite names, authors using a temporary foreign address, name mistakes, publications in SCI-journals, authors focusing on books and chapters, and fields with low citation and publication levels). The homonym problem would be much bigger if one uses also the full science citation index and papers with non-Dutch addresses. To reduce that risk, the search strategy was restricted as described, with the price of underestimating performance in several cases. As said no field normalization was done due to a lack of data access.

RESEARCH QUESTION

The basic assumption was, however, that the error would occur for both the successful and the nonsuccessful applicants. In other words, the resulting error was expected to be randomly distributed over the applicants and not be correlated with the variables in the study: level of performance and grant success. In that case, the error would not influence the findings. However, it remains important to investigate to what extent the data collection procedure may have influenced the findings. This is not only relevant for the specific study but also for scientometrics studies in general: how far does one need to go in data cleaning and disambiguation, given the research question to be answered? What are the data requirements, given the type of study one wants to do? These are the questions we address in this paper.

DATA AND METHODS

We did a replication of the original study, with a few extensions: the data were collected and cleaned again;

we used the same indicators as in the original study, but additionally a series of field normalized indicators. After showing that the results are similar, we analyze the differences between the original and the new dataset. Together this enables us to answer the research questions.

Data

Information about the applicants was retrieved from the web, consisting of CVs and of home pages. Using this, we recollected the data manually from the WoS, and now included SCI-journals and foreign addresses. A manual crosschecked name-disambiguation was done using the CV information found on the web, resulting in a as good as possible clean data set.

For a subset (economics, psychology, and behavioral science), we downloaded the records from the WoS using the names and name variants found in the manual data collection. This set of downloads was disambiguated using the BMX tool,[8] and again semi-automatically cleaned and disambiguated and compared with the manually collected data. For this subset of applicants in the sample, we also calculated field-normalized performance scores. Table 1 summarizes the characteristics of the resulting data sets.

Methods

We firstly calculate the original indicators using three different datasets and then compare the findings. Then, we will also calculate a series of more advanced bibliometric indicators of the three mentioned fields (psychology, behavior, and economics) and compare the results again. After having done this, we calculate the error in the data: the differences in publication and citation counts between the new correct dataset and the original dataset. This we do for the successful and nonsuccessful applicants separately. Then, the error distributions are compared in order to find out whether they differ or not between the two groups.

Findings

Comparing indicators based on the original and the new dataset

Using the old and the new data, we recalculated the indicators: publications in the 3 years before the application; the number of citations received until December 31, 2006. In contrast to the 2009 study,[7] we here not only compare averages but also, as the data are rather skewed, compare the medians. Table 2 shows the results for the sample of 905 applications, for the original 2007 data, and the new manually collected and corrected data. We show the mean and median of the publications and citations for the successful applicants (S) and for an equally large group of BPNS. (We do not give the figures for the set of all rejected applicants, as they behave as expected: always lower than the successful applicants.) We also show the ratio of the scores of the successful and the BPNS applicants (S/BPNS). In the original data, the successful applicants receive on average slightly more citations than the best performing nonsuccessful applicants. However, they have a slightly lower average number of publications. With the new data, the pattern is the same, and the same holds for the conclusion that the successful applicants do not outperform the best nonsuccessful. Obviously, the new data support the findings from the 2009 paper, and the assumption that the error is evenly distributed over the

Table 1: Datasets used in this paper

Dataset	Data collection procedure	Domain and sample
Original (2007)	Automatic coupled No disambiguation No field normalization	Dutch address Social Science Citation Index n=905* All age groups All social sciences, included psychology, behavior, and economics
New (2014)	Manual retrieved Manual disambiguation Double checked No field normalization	Also international addresses Social Science Citation and Science Citation Index Expanded n=905* (out of 1100) All age groups All social sciences, included psychology, behavior, and economics
Normalized (2014)	Automatic coupled Semi-automatic disambiguation plus double manual control Double manual check of first initial With field normalization	Also international address Social Science Citation Index and Science Citation Index Expanded n=260** (early career researchers) Psychology, behavior, and economics

*905 applications=864 unique applicants; the original study included four funding schemes, but here we include for resource reasons only three, **260 applications=246 unique applicants.

compared groups therefore seems correct. When looking at the median scores, the BPNS applicants score better, suggesting that among the successful applicants there is a tail of relatively low performing.

Does field normalization make difference: a test on a subset

We also tested whether field normalization changes the results. For practical (time resource) reasons, we do this for a subset of applicants only (the early career grant applicants). For theoretical reasons, we restrict this part of the analysis to fields where international journals are the main publication media: economics, behavior and education, and psychology. For this subset and each of the datasets, we calculate publications and citation scores, as well as the field-normalized scores [Table 3]. Best performing nonsuccessful applicants are in all cases defined in terms of the (field normalized) citation counts for that specific dataset. The S/BPNS column shows the ratio between mean and median for the two relevant groups.

The 2007 data suggest that the best performing NS applicants are on average slightly better than the selected applicants and in terms of the medians much better. With the manually corrected 2014 data, the pattern has slightly changed. On average, the BPNS applicants publish marginally more but are on average slightly less cited. In terms of the medians, the best nonsuccessful still are performing better than the successful applicants, although the differences are somewhat smaller than in the 2007 data. But the overall picture remains the same.

In addition, calculated the field-normalized scores without (NCSf) and with a 2-year citation window (NCSf2Y). We calculated the scores for this subsample using all the three datasets. In this subsample, we again find that the BPNS applicants score better in almost all the citation-related indicators. The main difference is in the publication measurement. In the corrected field-normalized data, the successful applicants score on average higher in terms of publications. However, not focusing on averages but on medians, which is more adequate given the skewed distributions, shows that also in publications the BPNS applicants do overall perform at least equally to the successful ones. By including also the science citation index data, some high publishing applicants were identified that had substantially lower publication records when only including the social science citation index. Although these applicants influence the means, they do not very much influence the medians. Again, the BPNS applicants score at least equally high as the successful ones. The differences between the two groups are also in the same order of magnitude as in the original set, so in this specific case, leaving out field normalization had no influence our findings.

Table 2: Performance of successful and pest performing nonsuccessful grant applicants*: All social science fields, all age groups

	Mean			Median		
	S	BPNS	S/BPNS[#]	S	BPNS	S/BPNS[#]
Original data (2007*/**)						
Publications	4.6	5.8	0.8	3	4	0.8
Citations	37	33	1.1	9	16	0.6
New data (2014**)						
Publications	5.7	7.5	0.8	3	4	0.8
Citations	48	49	1.0	18	22	0.8

*905 cases, 864 unique applicants. The results for the all 642 nonsuccessful applicants are not shown. **The numbers are different from those in (van den Besselaar *et al*. 2009); we here analyze a (large) sample of the original dataset. [#]If the value in the cell is smaller than 1, the BPNS applicants perform better than the successful applicants. S=Successful applicants (*n*=223), BPNS=Best performing (citations) nonsuccessful applicants (*n*=223).

Table 3: Performance of successful and best performing nonsuccessful grant applicants*: Early career researchers in psychology, behavior, and economics

	Mean			Median		
	S	BPNS[##]	S/BPNS[#]	S	BPNS[##]	S/BPNS[#]
Original data 2007						
Publications	3	4	0.8	2.0	4.0	0.5
Citations	24	32	0.8	10	21	0.5
New data 2014						
Publications	4.8	4.9	1.0	3	4	0.8
Citations	41	34	1.2	18	24	0.8
Field normalized data 2014						
Publications	4.8	3	1.6	3	3	1
Publications (frac.)	1.6	1	1.6	1.3	0.9	1.4
NCSSC	1.5	2.3	0.7	1.2	1.9	0.6
NCSSC (2Y)	1.3	2.3	0.6	1.0	1.6	0.6
Top 1%	0.02	0.03	0.7	0	0	1
Top 5%	0.10	0.19	0.5	0	0	1
Top 10%	0.17	0.33	0.5	0	0.23	0
Top 25%	0.37	0.64	0.5	0.30	0.61	0.5
Top 50%	0.70	0.85	0.8	0.78	1.00	0.8
S > PNS (%)			22			11
S=BPNS (%)			0			33
S < BPNS (%)			78			55

*246 cases. The results for the set of all 196 nonsuccessful applicants are not included in this table, [#]If the cell value<1, BPNS applicants perform better than the successful applicants, [##]2007 original data= Best performance in terms of citations, [##]2014 new data=Best performance in terms of citations, [##]2014 normalized data=Best performance in terms of publications NCSSC (2Y). NCSSC=Field normalized citations, NCSSC (2Y)=Field normalized citations, two years citation window, Top X%=Share in top X% cited papers. S=Successful applicants (*n*=50), BPNS=Best performing nonsuccessful applicants (*n*=50).

Analyzing the error

The initial assumption that the measurement errors in the original data collection would be evenly distributed over the contrasted groups indeed seems correct and it does

not influence the statistical relationship between performance data and funding data. This can also be shown directly. After collecting the new data, the error in a number of publications and citations was calculated. In the Figures 1 and 2, we show the error distribution of the publications, for successful and nonsuccessful applicants. The mean error is about one in both groups, which means that on average we have missed one publication per applicant. For both groups, the standard deviation is about 3.5. Moreover, for both groups, the error distribution has a similar positive skew. Overall, the distribution of the error in the two groups is about the same and this again explains why we find the same results for the different datasets.

There are a few observations with a rather large error. In both groups, we have some highly overestimated applicants, but especially in the successful group, we also have a few strongly underestimated applicants. These observations may not influence the medians and even not the means.

However, the few underestimated applicants are in the corrected data in the top of the distributions, especially when we have rather small samples. Hence, with data containing substantial error, one should remain careful with conclusions about the top of the distribution. This indeed showed to be the case. Another study using the same datasets focused on gender differences in scholarly performance. The study concluded that (i) the gendered performance differences were disappearing and that (ii) if there is a difference, women seem to perform better.[9] The first conclusion remains when using the new data; however, the second not, as it was based exactly on the patterns in the top of the performance: with the original dataset, slightly more women were in the top of the performance distribution, but this was not the case anymore with the corrected data (note 5).[10]

We also show the graphs for the error in the citation counts [Figures 3 and 4]. Here, we find a similar pattern as with publications: a few overestimated applicants and a few strongly underestimated (successful) applicants.

Implications

What does this imply? The results support that the required level of data cleaning and disambiguation indeed relates to the questions to be answered. For example, if the noncleaned data contain many researchers with only a very few publications with no or a few citations, one may decide not to disambiguate those, as they any how are only noise. The signal is only in the high publishing authors and in publications that attract many publications. Disambiguation may concentrate on these relatively few authors

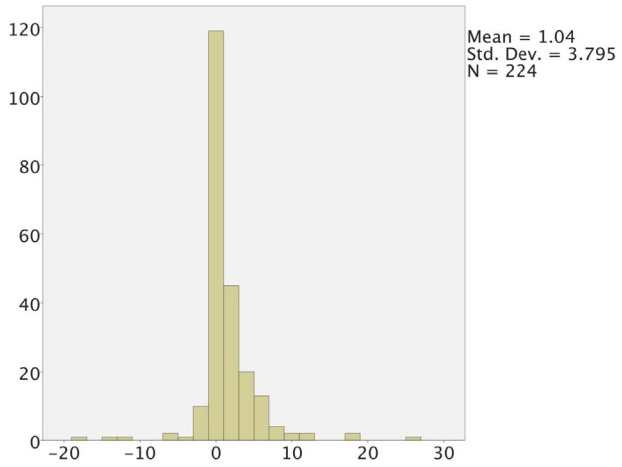

Figure 1: New counts of publications minus original counts successful applicants

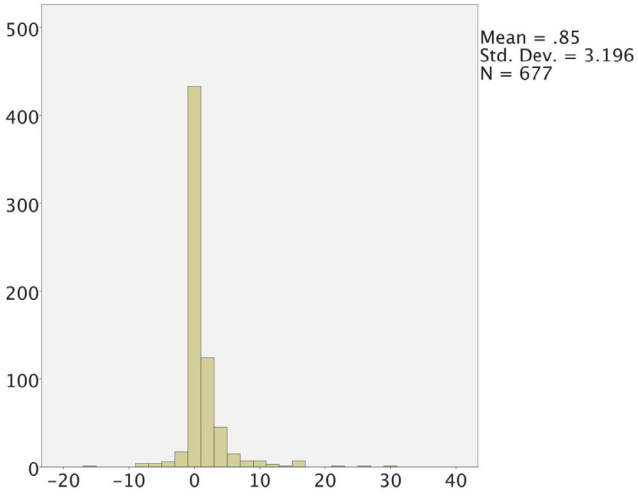

Figure 2: New counts of publications minus original counts nonsuccessful applicants

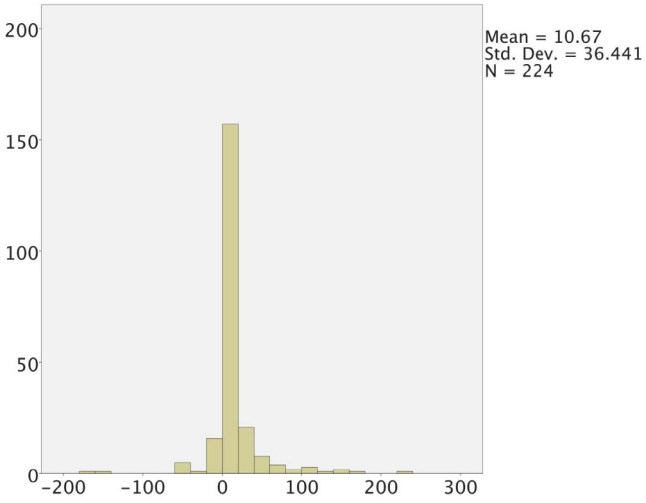

Figure 3: New counts of citations minus original counts successful applicants

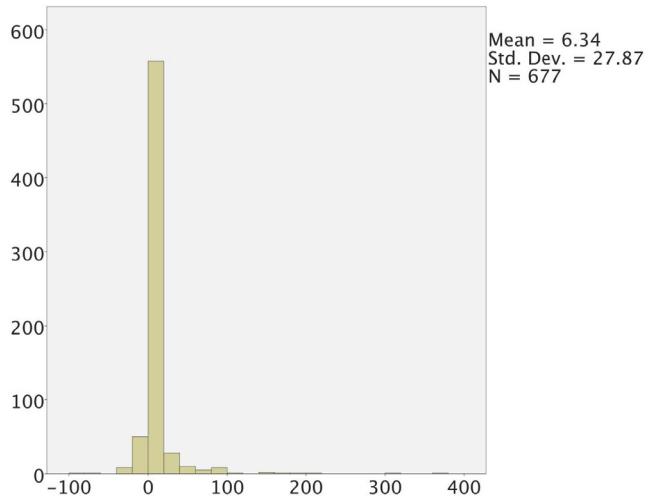

Figure 4: New counts of citations minus original counts non-successful applicants

and papers and that may reduce the required manual effort considerably. This, of course, will work easier for cleaning synonyms than cleaning homonyms. In the latter case, one may be faced with authors who use different first names or initials and this may influence recall: those other first names or initials may lead to not having part of the papers in the initial set to be cleaned. Getting the recall correct, before improving precision, is another problem that also asks for other approaches.

This all leaves open that there are also many types of research projects where a high level of cleaning is needed, for example, when network links are important or in applied research focusing on the evaluation of (small) groups and individuals. Then, one cannot accept much error in the data as these may have a big effect on the outcomes. Further research should try to develop instruments helping to decide about required levels of data cleaning, given the aim of research projects and the size of the samples.

Acknowledgments

The authors thank Charlie S. Mom and Pleun van Arensbergen for their efforts in the data collection and cleaning. We acknowledge the support of the RISIS project (EC grant 313082): Research Infrastructure for Science and Innovation Studies, especially task WP25 on disambiguation.

REFERENCES

1. Gurney T, Horlings E, van den Besselaar P. Author disambiguation using multi-aspect similarity indicators. Scientometrics. 2012;91:435-49.
2. Milojević S. Accuracy of simple, initials-based methods for author name disambiguation. J Inform. 2013;7:767-73.
3. Reijnhoudt L, Costas R, Noyons E, Börner K, Scharnhorst A. Seed expand: a general methodology for detecting publication oeuvres of individual researchers. Scientometrics. 2014;101:1403-17.
4. Newman ME. The structure of scientific collaboration networks. Proc Natl Acad Sci USA. 2001;98:404-9.
5. Kim JS, Diesner J. Distortive effects of initial-based name disambiguation on measurements of large-scale coauthorship networks. JASIST 2015; Early View. DOI: 10.1002/asi.23489.
6. Fegley BD, Torvik VI. Has large-scale named-entity network analysis been resting on a flawed assumption? PLoS One. 2013;8:e70299.
7. Van den Besselaar P, Leydesdorff L. Past performance, peer review, and project selection: A case study in the social and behavioral sciences. Res Eval. 2009;18:273-88.
8. Sandström U, Sandström E. The field factor: Towards a metric for academic institutions. Res Eval. 2009;18:243-50.
9. van Arensbergen P, van der Weijden I, van den Besselaar P. Gender differences in scientific productivity: a persisting phenomenon?. Scientometrics. 2012;93:857-68.
10. Van den Besselaar P, Sandström U. Gender differences in research performance and its impact on careers: a longitudinal case study. Scientometrics. 2016;106:143-62.

How to cite this article: Besselaar P, Sandström U. What is the required level of data cleaning? A research evaluation case. J Scientometric Res. 2016;5(1):7-12. Full text available at http://www.jscires.org/v5/i1