Journal of Informetrics 7 (2013) 722–729

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

The problem of citation impact assessments for recent publication years in institutional evaluations

Lutz Bornmann*

Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

ARTICLE INFO

Article history:
Received 28 January 2013
Received in revised form 15 May 2013
Accepted 15 May 2013
Available online 19 June 2013

Keywords:
Bibliometrics
Percentiles
Citation window
Population
Sample
Counterfactual concept of causality

ABSTRACT

Bibliometrics has become an indispensable tool in the evaluation of institutions (in the natural and life sciences). An evaluation report without bibliometric data has become a rarity. However, evaluations are often required to measure the citation impact of publications in very recent years in particular. As a citation analysis is only meaningful for publications for which a citation window of at least three years is guaranteed, very recent years cannot (should not) be included in the analysis. This study presents various options for dealing with this problem in statistical analysis. The publications from two universities from 2000 to 2011 are used as a sample dataset (n = 2652, univ 1 = 1484 and univ 2 = 1168). One option is to show the citation impact data (percentiles) in a graphic and to use a line for percentiles regressed on 'distant' publication years (with confidence interval) showing the trend for the 'very recent' publication years. Another way of dealing with the problem is to work with the concept of samples and populations. The third option (very related to the second) is the application of the counterfactual concept of causality.

1. Introduction

Modern science evaluates and is also subject to evaluation. Without research assessments, it is impossible to ensure the quality of research. That is why, according to the founder of the modern sociology of science Robert K. Merton (1973), one of its norms is "organised scepticism". From the 17th century, peer review was used almost exclusively to evaluate research until the 1980s and 1990s when indicator-based evaluation and multi-stage evaluation procedures were introduced (Daniel, Mittag, & Bornmann, 2007). It is now standard for an evaluation report of an institution to include bibliometric indicators on the number of publications and the citation impact of these publications (for the natural and life sciences). Appropriate standards such as those formulated by Bornmann et al. (in press) can be used to conduct a bibliometric study.

However, institutional evaluations frequently present the problem that it is precisely the research performance over very recent years that needs to be measured as interest is focussed on these years. It is only possible to measure the citation impact of a publication reliably around three years after it has appeared. The most recent 1 to 2 publication years of an institution cannot be included in the evaluation, even if methods of field normalization are used (Wang, 2013). According to the Council of Canadian Academies (2012) "past research suggested that, for the natural sciences and engineering, an appropriate citation window is typically between three and five years ... More recent evidence, however, has proposed that a citation window as short as two years may be appropriate in some cases ... This evidence implies that citation-based indicators should be limited to assessing research published at least two years previously. Any attempt to use citation-based

* Tel.: +49 89 2108 1265.
E-mail address: bornmann@gv.mpg.de

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.05.002

indicators for more recent research may result in spurious or misleading findings" (p. 68). This study therefore describes options for statistical procedures which allow a statement to be made about very recent publication years on the basis of those publication years which can be included in the evaluation (that is, earlier publication years). This study follows up on activities which Bornmann and Mutz (2013) initiated with their publication on the use of samples in institutional evaluations.

A number of advanced indicators are used in bibliometrics with which it is possible to measure the citation impact of publications from a research institution. They are used to show the citation impact achieved by a publication relative to the impact which other publications from the same year and in the same field have made (Rehn, Kronman, & Wadskog, 2007). Up to now, these indicators have been calculated by determining the average citation impact over the publications in a year and a field, but recently percentiles have been proposed as an important alternative (Bornmann & Mutz, 2011). A percentile is a value below which a certain proportion of publications fall: The higher the percentile for a publication, the more citations it has received compared to publications in the same field and publication year (Bornmann, Mutz, Marx, Schier, & Daniel, 2011). Although there is still some uncertainty concerning the exact method of calculating percentiles (Bornmann, in press; Bornmann, Leydesdorff, & Mutz, 2013), compared to earlier indicators they have the advantage that they do not require a (arithmetic) mean to be established. As distributions of citations are skewed to the right, the mean is not suitable as a measure of the central tendency. Percentiles are therefore used as an indicator of citation impact in this study.

2. Methods

In this study, the publications from two universities from 2000 to 2011 are used as a sample dataset. For each publication, the citation window extends from the publication to the end of 2011. There are total of 2652 publications (articles and reviews) for the universities (univ 1 = 1484, univ 2 = 1168); they published an average of 221 publications per year (univ 1 = 124, univ 2 = 97). The percentiles for the publications are researched in InCites. InCites (http://incites.thomsonreuters.com/) is a web-based research evaluation tool allowing the assessment of the productivity and citation impact of institutions. Percentiles are defined by Thomson Reuters as follows: "The percentile in which the paper ranks in its category and database year [that means, in its reference set], based on total citations received by the paper. The higher the number [of] citations, the smaller the percentile number. The maximum percentile value is 100, indicating 0 cites received. Only article types *article, note, and review* are used to determine the percentile distribution, and only those same article types receive a percentile value. If a journal is classified into more than one subject area, the percentile is based on the subject area in which the paper performs best, i.e. the lowest value" (http://incites.isiknowledge.com/common/help/h_glossary.html). InCites defines percentiles in the inverse direction than the standards in the literature (Bornmann & Marx, 2013).

In general, three steps are needed in order to calculate the percentiles for a reference set and all these steps can be differently conducted (Bornmann et al., 2013).

First, the rank-frequency function (see Egghe & Rousseau, 2006) is calculated. All publications in the set are ranked in decreasing order by their number of citations, and the number of publications in the (reference) set is determined.

Secondly, the minimum or maximum, respectively, of the percentile scale must be determined. InCites assign publications with 0 citations a percentile of 100. Furthermore, publications with a high citation impact are assigned a low percentile and publications with a low citation impact are assigned a high percentile in InCites. By assigning the value 100 to the publications with 0 citations it is ensured that the missing citation impact of publications is reflected in the percentiles in the same way in every case. Different values for publications with 0 citations would arise if percentiles are calculated without using a constant value of zero.

Thirdly, each publication is assigned a percentile based on the citation distribution (sorted in decreasing order). However, percentiles can be calculated in different ways (Cox, 2005). InCites and, for example, Rousseau (2012) calculate the quantiles – that is, the continuous variable from which percentiles can be derived by rounding – using the ranks (i) and the number of publications (n) ($i/n \times 100$). The formula $((i - 0.5)/n \times 100)$ derived by Hazen (1914) is used very frequently nowadays for the calculation of percentiles (for example by StataCorp, 2011).

The analyses for this study were performed with the statistical software Stata (StataCorp, 2011).

3. Results

Fig. 1 uses box plots to show the universities' distributions of the percentiles in each publication year. The recent publication years are also included in this figure. It is clearly visible for both universities that 2011 (on average) resulted in a significantly lower citation impact for the publications, compared to other years. Including the final year in statistical bibliometric analyses for an evaluation study or considering it in isolation would result in an erroneous representation of the performance of the two universities in terms of their citation impact. As a percentiles distribution such as that shown in Fig. 1 is not unusual, but can be seen generally in publication sets, recent years should not be included in an evaluation study and ways should be sought with which to achieve a generalising statement about the citation impact of a university based on the other years (which then relates to the recent years).

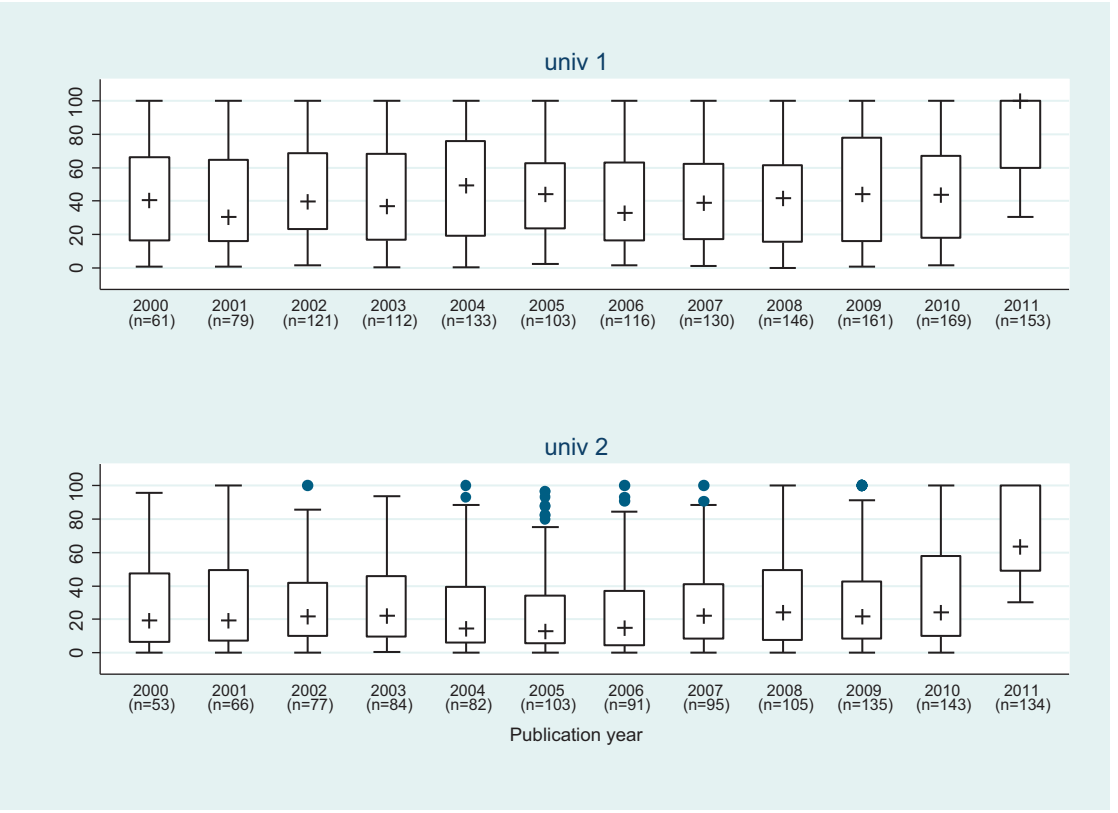

Fig. 1. Distribution of percentiles for the publications from two universities in individual publication years.

Three such options are presented in the following.

3.1. Regression fit for the percentiles of publications

The first option consists of representing the distribution of percentiles within the individual years with a scatter graph and integrating in this graphic a line for percentiles regressed on publication years (with confidence intervals) (Kohler & Kreuter, 2012). This line can then be used to estimate the citation impacts for two recent years (here: 2011 and 2012) based on the preceding years. Fig. 2 shows this regression fit for the percentiles of publications published by the universities. As the fitted lines shows the trend in the mean percentiles is slightly rising (the average citation impact has therefore fallen slightly over the years). We can therefore expect that for the two recent years, the mean percentiles will be slightly higher (i.e. a slightly lower citation impact) at around 46 for univ 1 and around 32 for univ 2. The slight rise which is visible for both universities in the Figure over all publication years might be a consequence of the shorter citation window for more recent years.

The first option "regression fit for the percentiles of publications" can be subsumed under the well-known idea of extrapolation. This idea can be applied not only to publications from recent years, but it can be applied equally well to publications that will appear in future years. For instance, in Fig. 2, extrapolation is done for the years 2011 and 2012, but in addition it could also be done for the years 2013 and 2014.

3.2. Complex samples (cluster samples)

A second way to determine the citation impact of recent years from that of earlier publication years is to view the publications from the earlier years as a sample from which to draw conclusions for the publications in the population. In this case, the population is made up of all the publications from one university (therefore including recently published publications), the percentiles of which can be researched in the InCites database (currently or – for the recent publication years – in the future). The sample forms a subset of the original set of measurements (the population) which is of interest to the evaluation (Levy & Lemeshow, 2008). In order to put together a sample (a publication subset) for a bibliometric study, the publications are not as a rule chosen at random from the literature database; certain publication years as non-overlapping clusters are selected (first step) and all the publications from these years (clusters) are collected into one sample (second step). According to Bornmann and Mutz (2013) this method of compiling the sample can be called a two-stage sampling

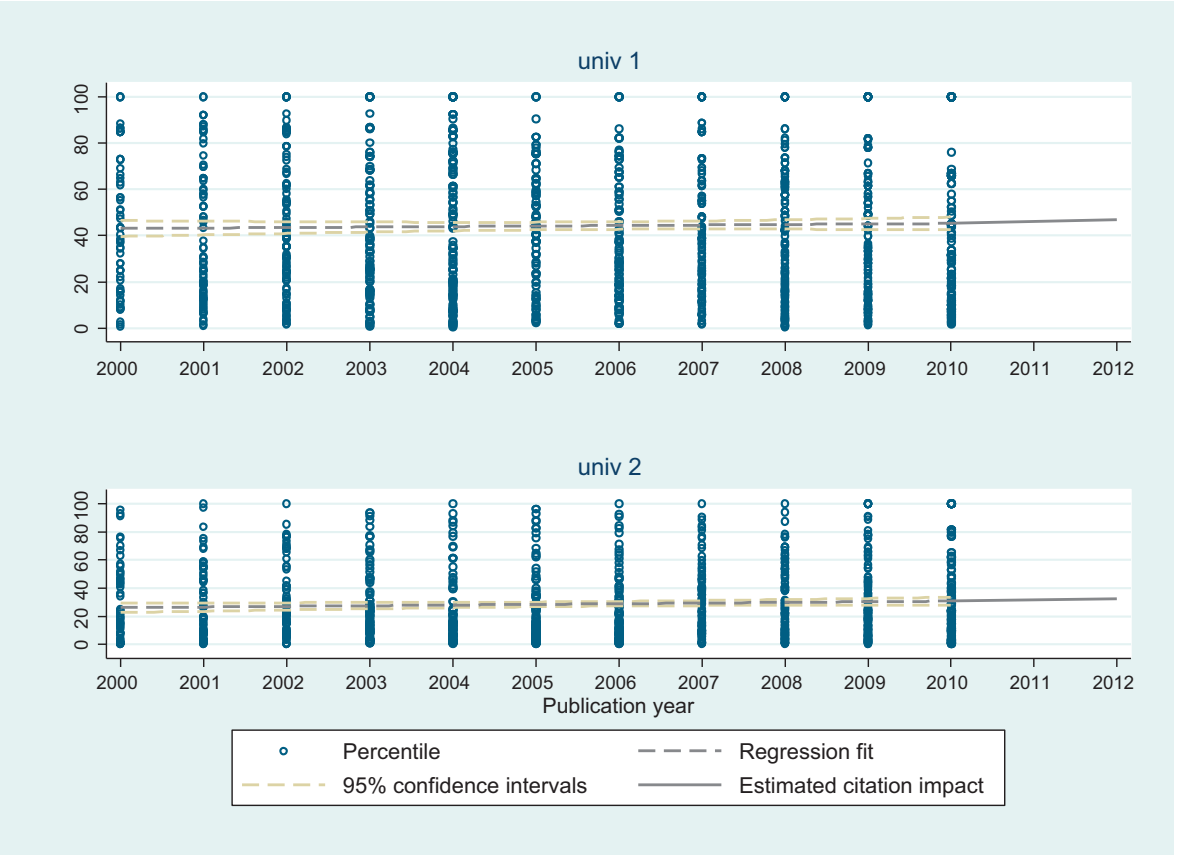

Fig. 2. Estimated citation impact for two recent publication years based on previous years for the publications of two universities.

design ("cluster sampling"). For the data example in this study, the publications from two universities from the years 2000 to 2010 were used and not the universities' publications for all years. The disadvantage of cluster samples compared to simple random samples[1] is that we can expect distortions from the use of clusters which affect the evaluation results. It is very likely that publications in the same cluster (here in the same publication year) are more similar than publications in different clusters.

Samples which are not simple random samples (here: cluster samples) are designated complex samples in Stata (Kohler & Kreuter, 2012). Because of the expected distortions (as described above) Stata offers a number of commands for dealing with complex samples (StataCorp, 2011). Stata's suite of complex sample data commands is governed by the 'svy' prefix. The prefix runs the supplied estimation command while accounting for the sample design characteristics in the point estimates and variance estimation method. In this study, the 'svy' prefix is used to compute standard errors by using the linearized variance estimator. This estimator is based on a first-order Taylor series linear approximation (Wolter, 2007). In addition to Stata, this kind of statistical procedures to analyse data from complex samples can also be undertaken by using (1) an add-on module for the SPSS package (http://www-142.ibm.com/software/products/us/en/spss-complex-samples/), (2) the 'survey' package in R (Lumley, 2010) or (3) certain SAS procedures (e.g., the SURVEYFREQ procedure) (http://support.sas.com/rnd/app/da/new/dasurvey.html).

The Stata commands for complex samples are usually applied in two stages: First a dataset is declared a complex sample. Then the relevant estimation command is used with the prefix 'svy'. As an example in this study, mean percentiles, corresponding standard errors and confidence intervals have been calculated for both universities twice: once taking the sample design (complex sample) into account and once ignoring it. The results are shown in Table 1. They indicate that the mean percentiles do not differ, but the standard errors and confidence intervals do. The 'design effect' can be calculated by dividing the standard error for the complex sample by the standard error for which the design was not taken into account. For example, this gives for univ 1 a value of 1.3, which indicates that the standard error of the cluster sample is 1.08/.83 = 1.3 times larger than the standard error when the cluster design is ignored. Both design effects with values larger than 1 in

[1] In statistics, a simple random sample is a subset from the elements in a population where each element has the same probability of selection (Levy & Lemeshow, 2008).

Table 1
Average percentiles with standard errors calculated by taking into account and ignoring the sample design ($n = 2365$).

	Mean percentile	Linearized standard error	95% confidence interval
Taking the sample design into account			
univ 1	44.33	1.08	41.92, 46.74
univ 2	28.84	1.18	26.20, 31.48
Ignoring the sample design			
univ 1	44.33	.83	42.71, 45.95
univ 2	28.84	.83	27.21, 30.47

Notes. The design effects are for univ 1 = 1.3 and univ 2 = 1.42.

Table 1 indicate that the cluster sample for the universities should be significantly larger than a simple random sample in order to achieve the same accuracy for the sample mean of the percentiles (Kohler & Kreuter, 2012).

If the correct standard error for a given sampling method (here: cluster sampling) can be estimated (by using the corresponding Stata commands), it is possible to calculate confidence intervals and significance tests as further uses of standard errors. With reference to the sample dataset used here, a significance test can be used to answer the following question: Do the two universities (univ 1 and univ 2) differ with respect to the citation impact of their publications randomly due to sampling variability or can it be safely assumed that there is also a difference in the population (and therefore for the recent publication years)? Table 1 shows the mean percentiles for the two universities. To account for the complex sample structure in significance testing, the Stata command 'test' after applying the 'mean' command can be used with the 'svy' prefix. 'test' performs an adjusted Wald test of a specified expression (Kohler & Kreuter, 2012; Sheskin, 2007). Here, it is tested whether the mean difference could be 0 in the population. As the results show, the null hypothesis can be rejected, $F(1, 11) = 160.95$, $p < 0.0001$: The difference is statistically significantly different from 0. We can therefore very confidently assume that there is a difference in the population and therefore also in the recent publication years.

3.3. Counterfactual concept of causality

The possibility in Stata (and in other statistical packages) of taking the sample design into account provides bibliometricians with very useful options for the statistical analysis of institutional publication sets, which are commonly based on cluster samples. However, there is a problem with the statistical analysis of cluster samples which is discussed in the final paragraph of this section: the cluster(s) with publications from certain years which are used to evaluate a university are usually not selected at random. They are usually deliberately chosen publication years which exclude very distant years. Even though these are not random samples (and statistical inference requires data drawn from the population by random sampling), we would nevertheless like to know from the statistical analysis whether the citation performance measured on the base of sample data could have happened by chance or whether there is a systematic (causal) relationship between university and citation impact. The concept of causal inference can be used to answer this question (Kohler & Kreuter, 2012).

In statistics, the counterfactual concept of causality is often used to answer questions concerning causality. It can be traced back to Rubin (1974) (Holland, 1986). According to Kohler and Kreuter (2012), the concept can be defined as follows: "A causal effect of some treatment T is the difference between an outcome Y^T of a specific research unit i if that unit experiences the treatment and the outcome Y^C of the same unit if that unit experiences the control conditions. Formally,

$$\delta_i = Y_i^T - Y_i^C$$

where δ_i is the causal effect" (p. 244).

When we relate the concept to the sample data in this study, we are investigating in the evaluation of the two universities (univ 1 and univ 2) in how far research demonstrates a systematically better or worse citation performance, because it has been done at univ 1 or univ 2. According to the counterfactual concept of causality, a systematically better or worse citation performance would be due to the university only if this difference were visible in research on the same topic – once with the univ 1 condition and once with the univ 2 condition. Simultaneous and independent research on the same topic at different universities is however a situation in the academic world which is difficult to identify.

Another point (in addition to research on the same topic) which should be taken into account in analyses using the counterfactual concept of causality is the comparability of research outcomes (of univ 1 and univ 2). As the success of universities is evaluated with citation impact, it should be ensured that the outcomes of the research – the publications – are comparable. As a number of studies has shown (Bornmann & Daniel, 2008; Bornmann, Schier, Marx, & Daniel, 2012), the citation impact of publications – apart from their quality – is dependent on many factors, such as the document type or the journal in which they appear. It is therefore necessary for analyses using the counterfactual concept of causality to select publications from the universities being examined which (i) have been produced as part of the research on the same topic and (ii) exhibit similarities in those factors which bibliometric studies have shown to have an effect on citation counts.

To illustrate the causality concept with the sample data set used in this study, it follows an analysis which relates to just one factor (the document type) where influence on citation impact can be assumed. The concept is therefore represented in a very simplified form as neither the many other factors that influence citation counts nor the restriction to research on the same topic are taken into account. The proportion of papers from both universities belonging to the 10% most cited papers

Table 2
Proportions, standard errors and confidence intervals of $PP_{top\ 10\%}$ for two universities ($n = 2215$ articles).

	Proportion	Standard error	95% confidence interval
univ 1	.14	.01	.12, .16
univ 2	.31	.02	.28, .34
Difference	−.17	.02	−.20, −.13

in their subject category is used in the analysis as an indicator of scientific success. This indicator is designated as $PP_{top\ 10\%}$ (Leydesdorff & Bornmann, 2012; Waltman et al., 2012) or as excellence rate (Bornmann, de Moya Anegón, & Leydesdorff, 2012). For univ 1 ($n = 1331$) $PP_{top\ 10\%}$ is 15% and 32% for univ 2 ($n = 1034$), so univ 2 achieves better results than univ 1. The two universities differ however in the document type of their publications. As univ 2 has published significantly more reviews (13%, $n = 130$) than univ 1 (2%, $n = 20$), the better citation performance of univ 2 could derive from the larger number of published reviews (only reviews and articles from the universities were included in this study, see above). A number of studies has already shown that reviews as a rule achieve a higher citation impact than publications with the 'article' document type (Bornmann & Daniel, 2008).

As the analysis of $PP_{top\ 10\%}$ only for the publications from both universities with the 'article' document type shows, the percentages (14% for univ 1 and 31% for univ 2) hardly differ from the percentages for all publications. Therefore, taking the document type into account in the analysis has hardly any effect. The performance difference remains and thus, seems to be systematic and stable. In this illustrative study only the document type is taken into account and it is assumed for the final stage of the statistical analysis that in the subset of publication data (the articles from both universities) there are hardly any other differences between the two universities (concerning the research topics and other factors influencing citation counts). In the following, therefore, we will proceed as if we have compiled a set with the selection of articles only with which both universities can be compared using the counterfactual concept of causality. Using this subset of publication data as a basis, we would now like to try to find an answer to the following inferential research question: How likely is it that the difference of 17 percentage points in $PP_{top\ 10\%}$ between both universities could arise if there were no systematic processes at stake (see Kohler & Kreuter, 2012).

If the publication data for both universities had really been created with a completely non-systematic process, then the difference in $PP_{top\ 10\%}$ should be within the limits that pure random fluctuation allows. To test this hypothesis, the standard error of a proportion can be estimated with the Stata 'proportion' command (here, R offers the prop.test() function). As the results of the estimation in Table 2 show, random processes would result in a $PP_{top\ 10\%}$ for univ 1 ranging from 12% to 16% and from 28% to 34% for univ 2. As the two intervals do not overlap, we can assume very safely that the observed difference between the universities cannot be attributed to random fluctuations. Using the Stata 'test' command (see above), we can estimate how confident we can be. As the result shows, the probability of observing a difference of 17 percentage points in $PP_{top\ 10\%}$ when there are only random processes running is 0, $F(1, 2214) = 83.58$, $p < 0.0001$. Correspondingly, the 95% confidence interval on the difference [−.20, −.13] misses zero.

4. Discussion

Bibliometrics has become an indispensable tool in the evaluation of universities (in the natural and life sciences). An evaluation report without bibliometric data has become a rarity – especially in the United Kingdom where "institutions are systematic about collecting metrics, including statistics on papers published and student-evaluation measures" (Abbott et al., 2010, p. 862). However, evaluations are often required to measure the citation impact of publications in very recent years in particular. As a citation analysis is only meaningful for publications for which a citation window of at least three years is guaranteed, very recent years cannot (should not) be included in the analysis (Wang, 2013). This study presents various options for dealing with this problem in statistical analysis. One option is to show the citation impact data (percentiles) in a graphic and to use a line for percentiles regressed on 'distant' publication years (with confidence interval) showing the trend for the 'very recent' publication years. In most cases we can assume that this trend line is a relatively reliable representation of the citation impact for the recent years, as the citation impact of a university as a rule does not change fundamentally. As more significant changes in citation impact are only expected when the publication pattern of a university changes (fundamentally) the publications in the very recent years should be compared to the publications from the earlier years: Did they appear in the same journals? Has the number of authors per publication, the average number of pages of the publications and other characteristics changed very much?

Another way of dealing with the problem of very recent publications is to work with the concept of samples and populations (Levy & Lemeshow, 2008). Other researchers, such as Hoffmann and Doucette (2012) also see working with samples in bibliometrics as an interesting option: "Rather than collecting, verifying, and analysing every cited reference, researchers should consider strategies for working with a more manageable number. Using a sample of citations is one approach ... Sampling is most effectively used when the total number of publications is large enough that a subset of those publications will still be representative". Bornmann and Mutz (2013) see the publication years selected for an evaluation study as a cluster sample, from which it is possible to draw conclusions about the population (all the publication years). The Stata program (and other statistical packages) offers a number of commands for analysing cluster samples which allow a reliable analysis.

For example, it is possible to use significance tests to determine in how far there will also be a citation impact difference between two universities in the population and therefore in the very recent publication years. It makes sense to use the sample to test also the practical significance besides the statistical significance of an outcome (Bornmann, 2013; Bornmann & Leydesdorff, 2013; Bornmann & Williams, 2013; Schneider, 2013). The practical significance provides information about the size of an effect, such as the size of a citation impact difference.

Including 'distant' publication years in an evaluation study is not only meaningful in terms of the inference on the very recent years. Their inclusion should also help to avoid the risk of a shifting baseline (Pauly, 1995). This is a risk posed by evaluations when a baseline used for the assessment of a certain outcome is not completely suitable, because it relates only to the most recent and not the more distant past. Pauly (1995) formulated this phenomenon for fisheries: "Essentially, this syndrome has arisen because each generation of fisheries scientists accepts as a baseline the stock size and species composition that occurred at the beginning of their careers, and uses this to evaluate changes. When the next generation starts its career, the stocks at that time serve as a new baseline." It also appears in research evaluation – against the background of the risk of a shifting baseline – to make sense not only to evaluate a small section (the most recent period) from many years of research at a university, but a period that is as long as possible. The performance of a university is thus seen from a long-term perspective and not as a snapshot, which would be susceptible to a shifting baseline.

The third option for dealing with the problem of very recent publication years presented in this paper is the application of the counterfactual concept of causality. In a comparison of two universities, one looks at in how far *similar* research undertaken at the universities demonstrates a better or worse performance. In this comparison, it should also be taken into account that not only the research of the universities is similar but also the publications in which the results are published. Using the "document type" example as one of the properties of publications which can be assumed to have an influence on citation impact, this study shows how an analysis against the background of the counterfactual concept of causality might look. As a rule, it would not be sufficient to take only one variable (e.g., the document type) into account in these analyses. As many variables as possible should be used to form subsets to be included in the analysis. However, it should be ensured that there are sufficient publications in a subset for the statistical analysis.

5. Conclusions

With the three options (regression fit, concept of samples and populations, and counterfactual concept of causality) presented in this paper, it should be possible to carry out a bibliometric study without including the very recent publication years and to make statements about the very recent years with the aid of the more 'distant' publication years.

References

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & van Noorden, R. (2010). Metrics: Do metrics matter? *Nature*, *465*, 860–862.

Bornmann, L. (2013). How to analyse percentile citation impact data meaningfully in bibliometrics: The statistical analysis of distributions, percentile rank classes and top-cited papers. *Journal of the American Society for Information Science and Technology*, *64*(3), 587–595.

Bornmann, L. (in press). Assigning publications to multiple subject categories for bibliometric analysis: An empirical case study based on percentiles. *Journal of Documentation* (in press)

Bornmann, L., Bowman, B. F., Bauer, J., Marx, W., Schier, H., & Palzenberger, M. (2013). Standards for using bibliometrics in the evaluation of research institutes. In B. Cronin, & C. Sugimoto (Eds.), *Next generation metrics*. Cambridge, MA, USA: MIT Press (in press).

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80. http://dx.doi.org/10.1108/00220410810844150

Bornmann, L., de Moya Anegón, F., & Leydesdorff, L. (2012). The new Excellence Indicator in the World Report of the SCImago Institutions Rankings 2011. *Journal of Informetrics*, *6*(2), 333–335. http://dx.doi.org/10.1016/j.joi.2011.11.006

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

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics*, *7*(1), 158–165.

Bornmann, L., & Marx, W. (2013). How good is research really? *EMBO Reports*, *14*(3), 226–230. http://dx.doi.org/10.1038/embor.2013.9

Bornmann, L., & Mutz, R. (2011). Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of Informetrics*, *5*(1), 228–230.

Bornmann, L., & Mutz, R. (2013). The advantage of the use of samples in evaluative bibliometric studies. *Journal of Informetrics*, *7*(1), 89–90. http://dx.doi.org/10.1016/j.joi.2012.08.002

Bornmann, L., Mutz, R., Marx, W., Schier, H., & Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: Do the editors of a high-profile journal select manuscripts that are highly cited after publication? *Journal of the Royal Statistical Society – Series A (Statistics in Society)*, *174*(4), 857–879. http://dx.doi.org/10.1111/j.1467-985X.2011.00689.x

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, *6*, 11–18.

Bornmann, L., & Williams, R. (2013). How to calculate the practical significance of citation impact differences? An empirical example from evaluative institutional bibliometrics using adjusted predictions and marginal effects. *Journal of Informetrics*, *7*(2), 562–574.

Council of Canadian Academies. (2012). *Informing research choices: Indicators and judgment: The expert panel on science performance and research funding*. Ottawa, Canada: Council of Canadian Academies.

Cox, N. J. (2005). Calculating percentile ranks or plotting positions. Retrieved from http://www.stata.com/support/faqs/stat/pcrank.html

Daniel, H.-D., Mittag, S., & Bornmann, L. (2007). The potential and problems of peer evaluation in higher education and research. In A. Cavalli (Ed.), *Quality assessment for higher education in Europe* (pp. 71–82). London, UK: Portland Press.

Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, *69*(1), 121–129.

Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of American Society of Civil Engineers*, *77*, 1539–1640.

Hoffmann, K., & Doucette, L. (2012). A review of citation analysis methodologies for collection management. *College & Research Libraries*, *73*(4), 321–335.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.
Kohler, U., & Kreuter, F. (2012). *Data analysis using Stata* (3rd ed., pp.). College Station, TX, USA: Stata Press, Stata Corporation.
Levy, P., & Lemeshow, S. (2008). *Sampling of population – Methods and applications* (4th ed., pp.). New York, NY, USA: Wiley.
Leydesdorff, L., & Bornmann, L. (2012). Testing differences statistically with the Leiden ranking. *Scientometrics*, *92*(3), 781–783.
Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Hoboken, NJ: Wiley.
Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago, IL, USA: University of Chicago Press.
Pauly, D. (1995). Anecdotes and the shifting baseline syndrome of fisheries. *Trends in Ecology & Evolution*, *10*(10), 430. http://dx.doi.org/10.1016/S0169-5347(00)89171-5
Rehn, C., Kronman, U., & Wadskog, D. (2007). *Bibliometric indicators – Definitions and usage at Karolinska Institutet*. Stickholm, Sweden: Karolinska Institutet University Library.
Rousseau, R. (2012). Basic properties of both percentile rank scores and the I3 indicator. *Journal of the American Society for Information Science and Technology*, *63*(2), 416–420. http://dx.doi.org/10.1002/asi.21684
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.
Schneider, J. W. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Informetrics*, *7*(1), 50–62. http://dx.doi.org/10.1016/j.joi.2012.08.005
Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed., pp.). Boca Raton, FL, USA: Chapman & Hall/CRC.
StataCorp. (2011). *Stata statistical software: Release 12*. College Station, TX, USA: Stata Corporation.
Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., et al. (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.
Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851–872. http://dx.doi.org/10.1007/s11192-012-0775-9
Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed., pp.). New York, NY, USA: Springer.