Qual Quant (2015) 49:155–165
DOI 10.1007/s11135-013-9979-1

Research quality evaluation: comparing citation counts considering bibliometric database errors

Fiorenzo Franceschini · Domenico Maisano · Luca Mastrogiacomo

Published online: 24 December 2013

Abstract When evaluating the research output of scientists, institutions or journals, different portfolios of publications are usually compared with each other. e.g., a typical problem is to select, between two scientists of interest, the one with the most cited portfolio. The total number of received citations is a very popular indicator, generally obtained by bibliometric databases. However, databases are not free from errors, which may affect the result of evaluations and comparisons; among these errors, one of the most significant is that of omitted citations. This paper presents a methodology for the pair-wise comparison of publication portfolios, which takes into account the database quality regarding omitted citations. In particular, it is defined a test for establishing if a citation count is (or not) significantly higher than one other. A statistical model for estimating the type-I error related to this test is also developed.

Keywords Database quality · Database error · Citation count · Omitted citations · Pair-wise comparison

1 Introduction and literature review

Bibliometric databases can be affected by different types of errors, whose consequences can be more or less severe (Kim et al. 2003). The information contained into databases is commonly used to (i) allocate resources between research institutions, (ii) support competitive academic examinations, (iii) drive subscriptions to scientific journals, etc. (Dalrymple et al. 1999; Adam 2002; Casser and Husson 2005; Guilera et al. 2010). Because of these important implications, the problem of database errors has been occasionally debated by the scientific community since the early years after the introduction of databases (Sweetland 1989; Abt 1992).

F. Franceschini (✉) · D. Maisano · L. Mastrogiacomo
Department of Management and Production Engineering (DIGEP), Politecnico di Torino,
Corso Duca degli Abruzzi 24, 10129 Turin, Italy
e-mail: fiorenzo.franceschini@polito.it

Databases have gradually improved the reliability of their contents thanks to the usual implementation of automatic tools for checking/correcting errors in the cited article lists (Adam 2002). The credit for this improvement should also be shared with reviewers, editors and publishers that—in recent years—have been paying more attention to the accuracy (quality) of published information, especially as regards bibliographic reference lists.

Despite recent improvements, the problem of bibliometric database errors is far from being solved. Several recent articles documented the existence of different types of errors (Buchanan 2006; Jacsó 2006; Li et al. 2010; Jacsó 2012). Among these errors, certainly one of the most significant is that of omitted citations, i.e., "citations that should be ascribed to a certain (cited) paper—being given by (citing) papers that are theoretically indexed by the database in use—but, for some reason, are lost" (Franceschini et al. 2013). Depending on the cause, Buchanan (2006) classifies these errors into two categories:

- Errors made by authors when creating the list of cited articles for their publication;
- Database mapping errors, i.e. failures to establish an electronic link between a cited article and the corresponding citing articles that can be attributed to a data-entry error.

Based on a limited number of journals, Buchanan (2006) asserts that omitted citations are likely to be around 5–10 % of the total number of "true" citations.

Franceschini et al. (2013) recently proposed a novel automated methodology for estimating the rate of omitted citations. This method is based on the comparison of "overlapping" citation statistics concerning the same set of papers of interest, but provided by two (or more) different databases. In the absence of an absolute reference—i.e., the "true" number of citations received by the paper(s) of interest, which can never be known exactly—this redundancy of information allows a reasonable estimate of the level of accuracy of a database with respect to one other. A first application example on a small sample consisting of three journals showed a significant omitted citation rate [in line with that one estimated by Buchanan (2006)], for both the two major bibliometric databases, i.e. Web of Science (WoS) and Scopus.

Apart from the preliminary empirical results, the study by Franceschini et al. (2013) opened the way for a new approach to analyze the accuracy of databases. Borrowing the general concept from metrology, the data contained in a database—as any measurement in general—can be affected by errors. According to this logic, levels of confidence may be associated with the inferences drawn from these data.

Regarding citation analysis, the knowledge of the database errors concerning citation count is crucial for two general problems:

- Comparing the number of citations with a specific threshold.
  Given a certain portfolio of publications, is the number of received citations greater than a reference value? This problem concerns those contexts in which the performance of a paper, an author or a journal is determined by the number of citations achieved with respect to specific reference values (Franceschini et al. 2012, 2013; MIUR 2012);
- Comparing two citation counts.
  Given two publication portfolios, when can we state that one is significantly better than the other one?

While the former problem has already been tackled (Franceschini et al. 2013), the latter deserves further investigation. The purpose of this paper is to go into the latter problem, defining a suitable statistical significance test.

The remainder of this paper is structured in five sections. Section 2 presents an introductory example for illustrating the research problem. Section 3 recalls a statistical model by

Table 1 Total citations concerning the publications of two fictitious authors (A and B)

	Case-I	Case-II
C_A	200	200
C_B	195	190

Franceschini et al. (2013), which allows the estimation of the "true" number of citations. Section 4 illustrates several alternative statistical models (one exact and two approximated) for tackling the problem of comparing two portfolios of publications, based on the citations they obtained. Section 5 shows an application example of the statistical models. The concluding section highlights the main implications, limitations and original contributions of this manuscript.

2 Introductory example and problem definition

When comparing two portfolios of publications, it is quite usual to ask which has received more citations (Bornmann et al. 2008; Jeong et al. 2009). The quickest answer is obtained by querying bibliometric databases. Even if this approach is widely adopted, it treats citation counts as exact numbers, with no error.

Table 1 contains the citation counts (C_A and C_B) relating to the portfolios of two fictitious scientists (A and B), according to a fictitious database. Two situations are presented: in case-I the two authors have a quite similar number of citations, while in case-II their difference is a bit larger.

Since citation counts returned by a database are potentially affected by errors, can we state that $C_A > C_B$ in both the cases?

Let suppose that $\hat{C}_A^*$ and $\hat{C}_B^*$ represent an estimate of the "true" number of citations (C_A^* and C_B^*) obtained by the publications of authors A and B respectively. The previous statement is meaningful when the following null hypothesis (H_0) is satisfied:

$$H_0 : \mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*) \tag{1}$$

being $\mathrm{E}(\hat{C}_A^*)$ and $\mathrm{E}(\hat{C}_B^*)$ the expected values of $\hat{C}_A^*$ and $\hat{C}_B^*$ respectively. Thus, a statistically sound hypothesis testing is needed. The rest of the paper aims at addressing this issue.

3 The reference statistical model

In a recent paper, Franceschini et al. (2013) proposed a method for estimating the percentage (p) of omitted citations in bibliometric databases. The study also introduced a statistical model depicting the distribution of $\hat{C}^*$, i.e. the estimate of the "true" number of citations for a set of publications.

As a first step, the model considers the case of a single paper. Neglecting "phantom" citations—i.e. citations erroneously attributed to the document (Buchanan 2006; Jacsó 2006; Li et al. 2010; Jacsó 2012)—the relationship between (1) the "true" citations received by the generic i-th paper (c_i^*, i.e., citations given by papers that are purportedly indexed by the bibliometric database in use), (2) the real citations returned by the database (c_i) and (3) the citations omitted (o_i) by the database in use is modelled by:

$$\hat{c}_i^* = c_i + o_i. \tag{2}$$

In the proposed model, c_i is treated as a known constant parameter related to the i-th paper. On the other hand, o_i is estimated on the basis of the database omitted-citation rate (p) and treated as a random variable. $\hat{c}_i^*$ is the modelled estimate of the unknown parameter c_i^*. Being $\hat{c}_i^*$ a function of o_i, it is treated as a random variable too.

The expected value and variance of $\hat{c}_i^*$ are respectively:

$$\mathrm{E}\left(\hat{c}_i^*\right) = c_i + \mathrm{E}\left(o_i\right), \tag{3}$$

$$\mathrm{V}(\hat{c}_i^*) = 0 + \mathrm{V}(o_i) = \mathrm{V}(o_i). \tag{4}$$

To estimate the expected value and variance of $\hat{c}_i^*$, the estimation of $\mathrm{E}(o_i)$ and $\mathrm{V}(o_i)$ is required. The variable o_i can be modelled by a binomial distribution. Given (1) a generic i-th paper with c_i^* "true" citations and (2) the omitted-citation rate (p) related to articles homologous to the one of interest, the database's probability of omitting o_i citations is:

$$P\left(o_i\right) = \binom{c_i^*}{o_i} p^{o_i}\left(1-p\right)^{c_i^*-o_i}. \tag{5}$$

Since c_i^* is unknown, it can be replaced by $\mathrm{E}\left(\hat{c}_i^*\right)$, i.e., the best estimate of $\hat{c}_i^*$:

$$P\left(o_i\right) = \binom{\mathrm{E}\left(\hat{c}_i^*\right)}{o_i} p^{o_i}\left(1-p\right)^{\mathrm{E}\left(\hat{c}_i^*\right)-o_i}. \tag{6}$$

The expected value and the variance of the (random) variable o_i are respectively:

$$\mathrm{E}(o_i) = \mathrm{E}\left(\hat{c}_i^*\right) \cdot p, \tag{7}$$

$$\mathrm{V}(o_i) = \mathrm{E}\left(\hat{c}_i^*\right) \cdot p \cdot (1-p). \tag{8}$$

Combining Eqs. 3 and 7, it follows that:

$$\mathrm{E}\left(\hat{c}_i^*\right) = c_i + \mathrm{E}\left(\hat{c}_i^*\right) \cdot p. \tag{9}$$

From which it is obtained that:

$$\mathrm{E}\left(\hat{c}_i^*\right) = \frac{c_i}{1-p}. \tag{10}$$

Combining Eq. 4 with Eqs. 8 and 10, it is obtained that:

$$\mathrm{V}(\hat{c}_i^*) = \mathrm{E}(\hat{c}_i^*) \cdot p \cdot (1-p) = c_i \cdot p. \tag{11}$$

Finally $\mathrm{E}(o_i)$ and $\mathrm{V}(o_i)$ can be obtained as functions of c_i. Combining Eq. 7 with Eq. 10 and Eq. 4 with Eq. 11 it follows that:

$$\mathrm{E}\left(o_i\right) = c_i \cdot \frac{p}{1-p}. \tag{12}$$

$$\mathrm{V}(o_i) = \mathrm{V}(\hat{c}_i^*) = c_i \cdot p. \tag{13}$$

Leaving the perspective of a single i-th paper, similar considerations may apply to sets of papers. Considering a generic set of P papers for which a database provides $C = \sum_{i=1}^{P} c_i$ total citations, the total number of omitted citations is:

$$O = \sum_{i=1}^{P} o_i. \tag{14}$$

Assuming statistical independence among the o_i values related to different papers of the sample, the expected value and the variance of O are:

$$\mathrm{E}(O) = \sum_{i=1}^{P} \mathrm{E}(o_i) = \sum_{i=1}^{P} c_i \cdot \frac{p}{1-p} = C \cdot \frac{p}{1-p} \tag{15}$$

$$\mathrm{V}(O) = \sum_{i=1}^{P} \mathrm{V}(o_i) = \sum_{i=1}^{P} c_i \cdot p = C \cdot p. \tag{16}$$

Equations 15 and 16 could also be obtained by applying the binomial probability distribution function to a group of (P) articles with $C^* = \sum_{i=1}^{P} c_i^*$ (unknown) total "true" citations. Precisely, for a group of papers with C^* total "true" citations, the database's probability of omitting O citations is:

$$P(O) = \binom{\mathrm{E}\left(\hat{C}^*\right)}{O} p^O (1-p)^{\mathrm{E}\left(\hat{C}^*\right)-O}, \tag{17}$$

being $\mathrm{E}\left(\hat{C}^*\right)$ the best estimate of the parameter $\hat{C}^* = C + O$. In other terms, O can be modelled by a binomial distribution:

$$O \sim \mathrm{B}[n, p], \tag{18}$$

being:

p the percentage of omitted citations (estimated by the proposed method);

$n = \mathrm{E}(C^*)$. In practice this value is rounded to the nearest integer.

Given p, n can be calculated as:

$$n = \mathrm{E}(\hat{C}^*) = \frac{C}{1-p}. \tag{19}$$

The expected value and variance of O stem from the definition of the binomial distribution:

$$\mathrm{E}(O) = np \quad \Rightarrow \quad \mathrm{E}(O) = C \cdot \frac{p}{1-p}, \tag{20}$$

$$\mathrm{V}(O) = np(1-p) \quad \Rightarrow \quad \mathrm{V}(O) = C \cdot p. \tag{21}$$

Since O follows a binomial distribution and $\hat{C}^* = C + O$, the probability density function of $\hat{C}^*$ can be seen as a right shift of that one of O. The size of the shift is equal to C (which is treated as a deterministic parameter), hence the expected value and variance of $\hat{C}^*$ are respectively:

$$\mathrm{E}\left(\hat{C}^*\right) = C + np \quad \Rightarrow \quad \mathrm{E}\left(\hat{C}^*\right) = \frac{C}{1-p} \tag{22}$$

$$\mathrm{V}\left(\hat{C}^*\right) = np(1-p) \quad \Rightarrow \quad \mathrm{V}\left(\hat{C}^*\right) = C \cdot p. \tag{23}$$

4 Comparing sets of publications

Referring to the hypothesis testing mentioned in Sect. 2, the type-I error (α) is the risk of rejecting the null hypothesis H_0 (Eq. 1), when it is true. This risk can be expressed as:

$$\alpha = \Pr\ (\text{reject } H_0 \mid H_0 \text{ is true}) = \Pr\left(\hat{C}_A^* \le \hat{C}_B^* \middle| \mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*)\right). \tag{24}$$

According to the statistical model in Sect. 3, it can be said that:

$$\hat{C}_A^* = C_A + O_A$$
$$\hat{C}_B^* = C_B + O_B \tag{25}$$

where O_A and O_B are binomially distributed as

$$O_A \sim \mathrm{B}\,[n_A, p_A] \quad \Rightarrow \quad \Pr(O_A = h) = \binom{n_A}{h} p_A^h\,(1 - p_A)^{n_A - h}$$
$$O_B \sim \mathrm{B}\,[n_B, p_B] \quad \Rightarrow \quad \Pr(O_B = k) = \binom{n_B}{k} p_B^k\,(1 - p_B)^{n_B - k} \tag{26}$$

and n_A, n_B are calculated as:

$$n_A = \mathrm{E}(\hat{C}_A^*) = \frac{C_A}{1 - p_A}$$
$$n_B = \mathrm{E}(\hat{C}_B^*) = \frac{C_B}{1 - p_B}. \tag{27}$$

Combining Eqs. 24 and 25, one obtains that:

$$\alpha = \Pr\left(C_A + O_A \le C_B + O_B \,\middle|\, \mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*)\right)$$
$$= \Pr\left(O_A - O_B \le C_B - C_A \,\middle|\, \mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*)\right). \tag{28}$$

Being a difference between two binomially distributed random variables, the probability density function of $O_A - O_B$ is given by (Box et al. 1978):

$$\Pr(O_A - O_B = j) = \sum_{i=j}^{\min(n_A, n_B + j)} \Pr(O_A = i) \cdot \Pr(O_B = i - j). \tag{29}$$

which is defined for $j \in [-n_B, n_A]$ and $O_A \in [0, n_A]$, $O_B \in [0, n_B]$. Equation 29 is valid under the reasonable assumption of independence between O_A and O_B.

Knowing the probability density function of $O_A - O_B$ (Eq. 29), it is possible to evaluate α. If performed without any approximation, this (exact) calculation is not very practical. For simplifying it, two possible approximations are:

1. Under certain conditions, the (binomial) distributions of O_A and O_B can be approximated with two Poisson distributions and therefore their difference $(O_A - O_B)$ will follow a Skellam (1946) distribution;
2. Assuming that the distributions of O_A and O_B are well approximated with two normal distributions, the difference $O_A - O_B$ will follow a normal distribution too.

The following sub-sections will discuss these two approximations individually and then compare their fit to the exact approach.

4.1 Poisson approximation

A generic binomial distribution $\mathrm{B}\,[n, p]$converges towards a Poisson distribution when the number of trials (n) goes to infinity while the product np remains fixed. The Poisson distribution with parameter $\lambda = np$ (i.e., $P\,[\lambda]$) can therefore be used as an approximation to the binomial distribution, provided that n is sufficiently large and p is sufficiently small. According to a general rule of thumb, this approximation is acceptable when $n \ge 20$ and $p \le 0.05$, or $n \geq 100$ and $np \leq 10$ (Box et al. 1978). In general, these conditions are satisfied in our problem, hence O_A and O_B can be assumed to be distributed as

$$O_A \sim \mathrm{P}\left[\lambda_A = n_A \cdot p_A\right] \quad \Rightarrow \quad \Pr(O_A = h) = e^{-n_A p_A} \frac{(n_A p_A)^h}{h!}$$
$$O_B \sim \mathrm{P}\left[\lambda_B = n_B \cdot p_B\right] \quad \Rightarrow \quad \Pr(O_B = k) = e^{-n_B p_B} \frac{(n_B p_B)^k}{k!}. \tag{30}$$

Under the hypothesis of independence between O_A and O_B, their difference is distributed according to a Skellam (1946) distribution:

$$\Pr(O_A - O_B = i) = e^{-(\lambda_A - \lambda_B)} \left(\frac{\lambda_A}{\lambda_B}\right)^{\frac{i}{2}} \mathrm{I}_{|i|}\left(2\sqrt{\lambda_A \lambda_B}\right). \tag{31}$$

where $\mathrm{I}_{|i|}$ is the modified Bessel function of the first kind. The expected value and the variance of O_A—O_B and their estimates (obtained from the estimates of n_A, n_B in Eq. 27) are respectively:

$$\mathrm{E}(O_A - O_B) = (\lambda_A - \lambda_B) = n_A p_A - n_B p_B \quad \Rightarrow \quad \mathrm{E}(O_A - O_B) = \frac{p_A C_A}{1 - p_A} - \frac{p_B C_B}{1 - p_B}$$
$$\mathrm{V}(O_A - O_B) = (\lambda_A + \lambda_B) = n_A p_A + n_B p_B \quad \Rightarrow \quad \mathrm{V}(O_A - O_B) = \frac{p_A C_A}{1 - p_A} + \frac{p_B C_B}{1 - p_B}. \tag{32}$$

As a consequence, the type-I error is estimated as:

$$\alpha = \Pr\left(\hat{C}_A^* \leq \hat{C}_B^* \middle| \mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*)\right) \approx F_S\left(C_B - C_A, \lambda_A = \frac{p_A C_A}{1 - p_A}, \lambda_B = \frac{p_B C_B}{1 - p_B}\right) \tag{33}$$

where F_S is the cumulative density function (CDF) of the generic random variable following a Skellam distribution with estimated parameters λ_A and λ_B.

4.2 Normal approximation

When n is large enough (at least around 20) and $\frac{1}{n+1} \leq p \leq \frac{n}{n+1}$, another reasonable approximation to $\mathrm{B}[n, p]$ is given by the normal distribution (Box et al. 1978; Montgomery 2009):

$$N\left[np, np(1 - p)\right]. \tag{34}$$

A rule of thumb to test when this approximation is appropriate is that np (if $p < 1/2$) or $n(1 - p)$ (if $p > 1/2$) must be ≥ 5 (Box et al. 1978; Montgomery 2009). This other condition is generally fulfilled in our problem. Hence we can assume:

$$O_A \sim N\left[n_A p_A, n_A p_A(1 - p_A)\right]$$
$$O_B \sim N\left[n_B p_B, n_B p_B(1 - p_B)\right]. \tag{35}$$

As a consequence—under the hypothesis of independence between O_A and O_B—the difference between O_A and O_B will be normally distributed with parameters:

$$\mu = \mathrm{E}(O_A - O_B) = n_A p_A - n_B p_B \quad \Rightarrow \quad \mu = \frac{p_A C_A}{1 - p_A} - \frac{p_B C_B}{1 - p_B}$$
$$\sigma^2 = \mathrm{V}(O_A - O_B) = n_A p_A(1 - p_A) + n_B p_B(1 - p_B) \quad \Rightarrow \quad \sigma^2 = p_A C_A + p_B C_B. \tag{36}$$

Then the α-risk estimate is given by:

$$\alpha = \Pr\left(\hat{C}_A^* \le \hat{C}_B^* \middle| \mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*)\right)$$
$$\approx \Phi\left(\frac{(C_B - C_A) - \hat{\mu}}{\hat{\sigma}}\right) = \Phi\left(\frac{(C_B - C_A) - \left(\frac{p_A C_A}{1-p_A} - \frac{p_B C_B}{1-p_B}\right)}{\sqrt{p_A C_A + p_B C_B}}\right) \quad (37)$$

where Φ is the CDF of the standardized normal random variable.

4.3 Further considerations

In absence of database errors, the "true" number of citations would correspond to the citation count provided by the database. According to the notation in use, this would entail that:

$$\begin{cases} \hat{C}_A^* = \mathrm{E}(\hat{C}_A^*) = C_A \\ \hat{C}_B^* = \mathrm{E}(\hat{C}_B^*) = C_B \end{cases}. \quad (38)$$

In such conditions the probability $\Pr(\hat{C}_A^* \le \hat{C}_B^*)$ would trivially be

$$\Pr(\hat{C}_A^* \le \hat{C}_B^*) = \begin{cases} \Pr(C_A \le C_B | C_A > C_B) = 0 \\ \Pr(C_A \le C_B | C_A \le C_B) = 1 \end{cases}. \quad (39)$$

Not surprisingly, the risk α would be 0.

On the other hand, in presence of omitted citations, $\Pr(\hat{C}_A^* \le \hat{C}_B^*) \in [0,1]$. As modelled in the previous sections, this probability depends on:

- the number of citations (C_A and C_B) indexed by the database;
- the percentage (p_A and p_B) of citations omitted by the database.

Let us now introduce a qualitative example to better understand the effect of omitted citations on $\Pr(\hat{C}_A^* \le \hat{C}_B^*)$. Given a fixed value of $C_A = C_{A,Fixed}$, the probability $\Pr(\hat{C}_A^* \le \hat{C}_B^*)$ can be expressed as a function of C_B (which is varied). Fig. 1 shows a qualitative 2D plot of $\Pr(\hat{C}_A^* \le \hat{C}_B^*)$, comparing the ideal case of absence of error ("step" curve) with the real case (continuous monotonically increasing curve).

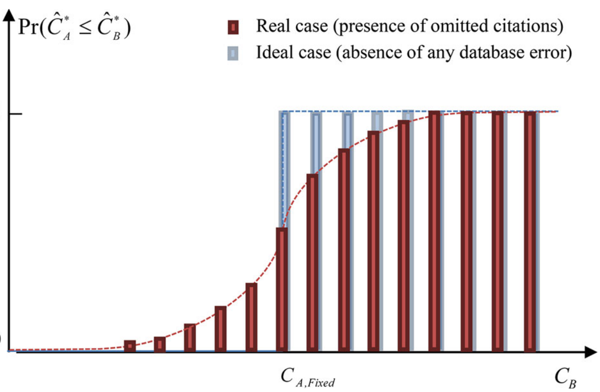

Fig. 1 Qualitative plot of $\Pr(\hat{C}_A^* \le \hat{C}_B^*)$: ideal versus real case

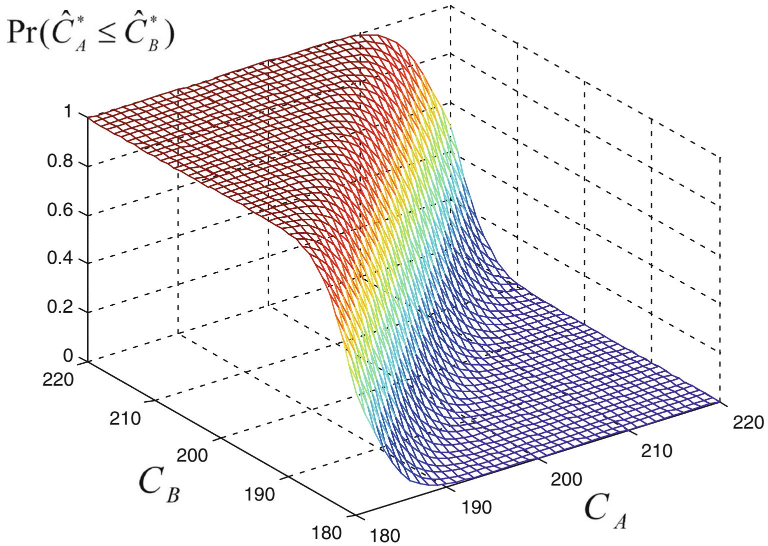

Fig. 2 Three-dimensional envelope of $\Pr(\hat{C}_A^* \leq \hat{C}_B^*)$, for values of C_A and C_B included in [180, 220] and $p_A = p_B = 5\,\%$

We remark that, in the real case, the value of $\Pr(\hat{C}_A^* \leq \hat{C}_B^*)$ tends to that one of the ideal case, as p decreases. Similar considerations hold when fixing C_B and varying C_A.

As a further quantitative example, the 3D plot in Fig. 2 shows the envelope of $\Pr(\hat{C}_A^* \leq \hat{C}_B^*)$, for values of C_A and C_B included in [180, 220] and $p_A = p_B = 5\,\%$.

5 Application example

Referring to the hypothesis testing in Sect. 2, let us now focus the attention on the type-I error (α).

Having defined a tolerable type-I error value (α_t), when

$$\alpha = \Pr(\hat{C}_A^* \leq \hat{C}_B^* \,\big|\, \mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*)) \leq \alpha_t, \tag{40}$$

it can be concluded that H_0 cannot be rejected. On the contrary, when

$$\alpha = \Pr(\hat{C}_A^* \leq \hat{C}_B^* \,\big|\, \mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*)) > \alpha_t \tag{41}$$

then H_0 is rejected.

For both the case-I and -II (in Table 1), we estimated the α values according to the three alternative approaches presented in Sect. 4 (see Table 2). It is assumed the independence between O_A and O_B, and a fraction of citations omitted by the database $p_A = p_B = 5\,\%$. Having defined a reasonable value of $\alpha_t = 0.05$, whatever the approach, in case-I the results suggest to reject H_0, being $\alpha > \alpha_t$. In other terms, when $C_A = 200$ and $C_B = 195$, the inference $\mathrm{E}(\hat{C}_A^*) > \mathrm{E}(\hat{C}_B^*)$ is too hasty and therefore not acceptable. On the contrary, in case-II there is no evidence to reject H_0, since $\alpha \leq \alpha_t$.

It is worth noticing that, in this specific example, the two approximations proposed in Sects. 4.1 and 4.2 (i.e., with the Skellam and Normal distribution respectively) work quite

Table 2 Type-I error (α) estimation according to the three approaches in Sect. 4

Approach	Case-I (%)	Case-II (%)
Exact evaluation	9.5	0.6
Skellam approximation	10.2	0.7
Normal approximation	11.8	0.8

It is assumed $p_A = p_B = 5\,\%$

well. We remark that the normal approximation is generally the simplest and consequently the most practical.

6 Conclusions

This paper proposed a statistical methodology for the pair-wise comparison of citation counts, which takes into account the omitted citation database errors.

In research quality evaluation this methodology may be of interest to database practitioners and users, as it establishes when a citation count is (or not) significantly higher than another, beyond any reasonable doubt. In some way, the proposed significance test represents a first attempt to deal with the uncertainty associated with the results obtained from bibliometric databases.

The main limitation is that, at the moment, this kind of testing only applies to pair-wise comparisons of citation counts. The subject of a future investigation will be extending these comparisons to more than two citation counts.

With some modifications, the suggested approach could be also extended to comparisons based on other bibliometric indicators, e.g., the total number of publications, the number of citations per paper (*CPP*), the h-index, etc. (Franceschini and Maisano 2010).

References

Abt, H.A.: What fraction of literature references are incorrect? Publ. Astron. Soc. Pac. **104**, 235–236 (1992)
Adam, D.: Citation analysis: the counting house. Nature **415**(6873), 726–729 (2002)
Bornmann, L., Mutz, R., Neuhaus, C., Daniel, H.D.: Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. Ethics Sci. Environ. Polit. **8**(1), 93–102 (2008)
Box, G.E.P., Hunter, W.G., Hunter, J.S.: Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building. Wiley, New York (1978)
Buchanan, R.A.: Accuracy of cited references: the role of citation databases. Coll. Res. Libr. **67**(4), 292–303 (2006)
Casser, D., Husson, F.: A 2-dimensional extension of the Bradley–Terry model for paired comparisons. J. Stat. Plan. Inference **135**(2), 245–259 (2005)
Dalrymple, J., Edgeman, R.L., Finster, M., Guerrero-Cusumano, J., Hensler, D.A., Parr, W.C.: A white paper: quality at the crossroads of organizational excellence and the academy. Qual. Eng. **12**(1), 97–104 (1999)
Franceschini, F., Galetto, M., Maisano, D., Mastrogiacomo, L.: The success-index: an alternative approach to the h-index for evaluating an individual's research output. Scientometrics **92**(3), 621–641 (2012)
Franceschini, F., Maisano, D.: Analysis of the Hirsch index's operational properties. Eur. J. Oper. Res. **203**(2), 494–504 (2010)
Franceschini, F., Maisano, D., Mastrogiacomo, L.: A novel approach for estimating the omitted-citation rate of bibliometric databases. J. Am. Soc. Inf. Sci. Technol. **64**(10), 2149–2156 (2013). doi:10.1002/asi.22898
Guilera, G., Gomez-Benito, J., Hidalgo, M.D.: Citation analysis in research on differential item functioning. Qual. Quant. **44**, 1249–1255 (2010)
Jacsó, P.: Deflated, inflated and phantom citation counts. Online Inf. Rev. **30**(3), 297–309 (2006)
Jacsó, P.: Grim Tales about the impact factor and the h-index in the Web of Science and the Journal Citation Reports databases: reflections on Vanclay's criticism. Scientometrics **92**(2), 325–354 (2012)

Jeong, S., Lee, S., Kim, H.G.: Are you an invited speaker? A bibliometric analysis of elite groups for scholarly events in bioinformatics. J. Am. Soc. Inf. Sci. Technol. **60**(6), 1118–1131 (2009)
Kim, W., Choi, B.J., Hong, E.K., Kim, S.K., Lee, D.: A taxonomy of dirty data. Data Min. Knowl. Discov. **7**(1), 81–99 (2003)
Li, J., Burnham, J.F., Lemley, T., Britton, R.M.: Citation analysis: comparison of Web of Science, Scopus, Scifinder, and Google Scholar. J. Electron. Resour. Med. Libr. **7**(3), 196–217 (2010)
MIUR: Decreto Ministeriale 7 giugno 2012 n. 76. http://attiministeriali.miur.it/anno-2012/giugno/dm-07062012.aspx (2012). Accessed 1 Oct 2013
Montgomery, D.C.: Statistical Quality Control: A Modern Introduction, 6th edn. Wiley, New York (2009)
Skellam, J.: The frequency distribution of the difference between two Poisson variates belonging to different populations. J. R. Stat. Soc. Ser. A (Gen.) **109**(3), 296 (1946)
Sweetland, J.H.: Errors in bibliographic citations: a continuing problem. Libr. Q. **59**(4), 291–304 (1989)